Les usages de l'intelligence artificielle

Edition 2021

Février 2021

Olivier Ezratty

A propos de l'auteur

Olivier Ezratty

consultant et auteur

olivier@oezratty.net, www.oezratty.net

+33 6 67 37 92 41, @olivez

Olivier Ezratty forme et conseille les entreprises dans l'élaboration de leurs stratégies d'innovation autour des *deep techs* et en particulier, des technologies quantiques et de l'intelligence artificielle. Il leur apporte une vision à 360° : scientifique, technologique, marketing ainsi que la connaissance des écosystèmes dans les industries numériques. Il a réalisé depuis 2005 des missions diverses d'accompagnement stratégique et de conférences ou formations dans différents secteurs tels que le secteur **médias/télécoms** (Orange, Bouygues Télécom, TDF, Médiamétrie, BVA), la **finance et l'assurance** (groupe BPCE, Caisse des Dépôts, Crédit Agricole, Crédit Mutuel-CIC, Generali, MAIF, Société Générale), l'**industrie et les services** (Schneider, Camfil, Vinci, NTN-STR, Econocom, ADP, Air France, Airbus) et le **secteur public** (Ministère des Armées, CEA, Météo France, Bpifrance, Business France). Ces missions couvrent des conférences, séminaires et formations ainsi que l'assistance à la définition de stratégies produits, d'innovation ouverte et de promotion d'écosystèmes.

Ses contributions s'appuient sur un fort investissement dans l'écosystème de l'innovation et sous différentes casquettes, notamment dans l'univers entrepreneurial :

- Formateur sur l'intelligence artificielle et l'informatique quantique auprès de **Cap Gemini Institut**.
- Membre depuis fin 2015 du Conseil Scientifique de l'**ARCEP**.
- Expert référent et intervenant à l'**IHEDN** dans les promotions 2019/2020 et 2020/2021.
- Expert auprès de **Wilco** – ex Scientipôle Initiative – depuis 2007, dans l'accélérateur santé depuis 2016.
- Membre du jury de divers **concours entrepreneuriaux** comme le Grand Prix de l'Innovation de la Ville de Paris ou le Concours National i-Lab et le Concours i-Nov pour le compte de Bpifrance.

Il intervient comme conférencier dans divers établissements d'enseignement supérieur tels que CentraleSupelec, l'Ecole des Mines de Paris, Télécom Paristech, l'EPITA, Les Gobelins, HEC, Neoma Rouen et SciencesPo, sur l'intelligence artificielle, les technologies quantiques ainsi que sur sur l'entrepreunariat et le product management, en français comme en anglais selon les besoins.

Olivier Ezratty est l'auteur des ebooks **Les usages de l'intelligence artificielle** en octobre 2017, novembre 2018, novembre 2019 et février 2021 et **Comprendre l'Informatique Quantique** en septembre 2018, 2019 et 2020, ainsi que du **Rapport du CES de Las Vegas**, publié en janvier chaque année entre 2006 et 2020 et du **Guide des Startups** mis à jour régulièrement depuis 2006 (23e édition et 13e année en 2019). Le tout étant publié sur le blog « Opinions Libres » (http://www.oezratty.net) qui traite de l'innovation technologique vue sous les angles scientifiques, technologiques, entrepreneuriaux et des politiques publiques de l'innovation. Comme photographe amateur, il est aussi le co-initiateur en 2012 de « Quelques Femmes du Numérique ! » devenu une association en 2016, et qui vise à augmenter la place des femmes dans les métiers du numérique, en sensibilisant les jeunes à ces métiers.

Et avant tout cela, Olivier Ezratty débute en 1985 chez **Sogitec**, une filiale du groupe Dassault, où il est successivement Ingénieur Logiciel, puis Responsable du Service Études dans la Division Communication. Il initialise des développements sous Windows 1.0 dans le domaine de l'informatique éditoriale ainsi que sur SGML, l'ancêtre de HTML et XML.

Entrant chez **Microsoft France** en 1990, il y acquiert une expérience dans de nombreux domaines du mix marketing : produits, canaux, marchés et communication. Il lance la première version de Visual Basic en 1991 ainsi que Windows NT en 1993. En 1998, il devient Directeur Marketing et Communication de Microsoft France et en 2001, de la Division Développeurs dont il assure la création en France pour y lancer notamment la plate-forme .NET et promouvoir la plate-forme de l'éditeur auprès des développeurs, dans l'enseignement supérieur et la recherche ainsi qu'auprès des startups.

Olivier Ezratty est ingénieur **Centrale Paris** (1985), devenu CentraleSupelec en 2015.

ISSN 2680-0527

photo de couverture : schéma trouvé ici et modifié.

Table des matières

Objectifs et contenu 5

Histoire et sémantique de l'IA 10
Hauts et bas de l'IA 12
Connexionisme et symbolisme 22
Définitions et segmentations 25
Limites de l'IA 29
Etat des lieux 31

Algorithmes et logiciels de l'IA 33
Force brute et arbres de recherche 35
Méthodes statistiques 36
Raisonnement automatique 37
Machine learning 45
Réseaux de neurones 54
Deep learning 59
Agents et réseaux d'agents 109
Artificial General Intelligence 112

Données de l'IA 141
Données d'entraînement 141
Données de test 142
Données de production 143
Données de renforcement 143
Origine des données 143
Biais des données 148
Audit de l'IA 155
Capteurs et objets connectés 155

Infrastructure de l'IA 159
Processeurs 160
Mémoire 223
Stockage 227
Big data 231
Cloud 241
AIOps 243
Robotic Process Automation 244
Energie 246
Cybersécurité 254

Applications génériques de l'IA 262
Vision 262
Langage 294
Robotique 327
Marketing et vente 342
Ressources humaines 352
Finance et achats 358

Applications métiers de l'IA 363
Transports 363
Bâtiments et Travaux Publics 390
Utilities 399

Industrie 404
Agro-alimentaire 411
Santé 424
Finance 466
Assurance 476
Juridique 481
Distribution 489
Télécoms 500
Médias et contenus 503
Services en ligne 529
Tourisme 530
Mode et luxe 534
Services et conseil 537
Education 540
Services publics 545
Renseignement et défense 549

Acteurs de l'IA 553
Grandes entreprises du numérique 553
Startups 596
Ecosystème français de l'IA 604

IA et société 611
Craintes sur l'IA 611
Emplois et IA 624
Politique par l'IA 648
Politiques de l'IA 652
Géopolitique de l'IA 679

IA et entreprise 707
Discours 707
Méthodes 710
Projets 714
Benchmarks 715
Outils 719
Financement 720
Compétences 720
Startups 723
Organisation 723
Ethique 724
Juridique 724

Epilogue 726

Médias spécialisés 727

Bibliographie 728

Dictionnaire anglais/français de l'IA 731

Glossaire 734

Historique des révisions du document 739

Objectifs et contenu

Cet ebook vous propose un tour d'horizon à 360° de la thématique de l'intelligence artificielle, pour en expliquer les origines, la démythifier, en décrire les principales techniques, pour en montrer la portée dans les solutions destinées aussi bien aux entreprises qu'au grand public, et pour traiter de ses retombées économiques, sociétales et politiques.

Alors que la notion même d'intelligence artificielle était presque bannie du langage commun depuis l'avènement du web en 1995, elle est redevenue au goût du jour au début des années 2010 et surtout aux alentours de 2015 avec l'émergence du deep learning dans les applications grand public. Comme de nombreuses thématiques liées au numérique, l'IA est aussi devenue un sujet d'attention par les questionnements qu'elle génère : sur le futur du travail, sur l'éthique, sur la répartition des richesses, sur la souveraineté économique et même sur le devenir de l'espèce humaine.

A l'instar de nombreuses technologies numériques, l'IA reste mystérieuse et prête le flanc à de nombreuses interprétations. Ses technologies et méthodes sont méconnues, y compris par la majorité des professionnels du numérique, d'où la propagation de nombreux mythes à son sujet, un peu trop directement inspirés par la science-fiction ou par les thèses singularistes[1].

On croit par exemple que le « deep learning » ou « apprentissage profond » raisonne alors qu'il ne fait qu'appliquer des modèles probabilistes. On pourrait aussi intuiter que l'apprentissage non-supervisé est de l'apprentissage supervisé « autonomisé » alors que ce n'est pas le cas. La sémantique de l'IA est porteuse de contre-sens qu'il faut décrire et éviter. Le sens même de l'appellation « IA » est sujet à d'infinis débats, connotés d'accents anthropomorphiques alors que l'IA est artificielle et donc différente de l'intelligence humaine, même si les deux peuvent présenter quelques zones de recouvrement. Et on peine encore à caractériser précisément les contours de l'intelligence humaine sans compter le fait qu'elle est inégalement répartie et, surtout, utilisée dans ce bas monde.

Grâce à une inexorable loi de Moore, qui n'est ni une loi mathématique ni inexorable ad-vitam en pratique, les singularistes voient arriver à grand pas une intelligence artificielle généralisée qui rendrait l'Homme caduque, la connexion directe des IA aux cerveaux et autres délires transhumanistes. Ces craintes ont pu saturer l'espace médiatique avec des propos qui se focalisent sur le pour et le contre de l'IA. L'éthique de l'IA et les biais de l'IA sont d'ailleurs plus connus que les techniques de l'IA ! L'IA est devenue le réceptacle et le révélateur de tous les défauts réels ou perçus du numérique et des logiciels.

Cet ouvrage se positionne depuis sa création à contre-courant de cette littérature anxyogène. Il vise à démythifier de manière posée l'état de l'art de l'IA et à rendre son adoption et son utilisation aussi pratiques que possible pour les entreprises. Ce document est aussi une grande boîte à idées pour les entreprises qui se demandent où et comment intégrer l'intelligence artificielle dans leur système d'information.

Cet ebook est le cinquième d'une série lancée en 2016 et dont voici les évolutions successives.

La **première édition**, Les avancées de l'intelligence artificielle, était une compilation de neuf articles publiés entre mars et mai 2016 sur mon blog Opinions Libres (197 pages). Elle débroussaillait à un premier niveau le brouhaha de l'époque sur l'IA. J'y décrivais la manière dont l'écosystème de l'IA se construisait à l'époque. Le contenu technique y était encore assez approximatif. Je m'intéressais notamment au cas d'IBM Watson qui me semblait survendu à l'époque.

[1] Mythes que j'ai eu l'occasion de décrire en septembre 2017 dans Douze mythes de l'intelligence artificielle. Voir aussi Mythes et légendes de l'intelligence artificielle par Clément Jeanneau, mai 2019 qui en reprend quelques autres (8 en tout).

La **seconde édition** Les usages de l'intelligence artificielle d'octobre 2017 (362 pages) complétait la précédente avec un contenu technique plus dense et précis, corrigeant au passage un bon nombre d'erreurs ou d'appréciations de la première édition. On y trouvait une partie dense sur le matériel de l'IA. Cette édition décrivait aussi les usages de l'IA par secteur d'activité, avec une douzaine de marchés comme les transports, la santé, la finance et la distribution.

La **troisième édition** de novembre 2018 était une très importante mise à jour de la seconde (520 pages). Les entreprises commençaient à prendre en main l'IA. Les études de cas étaient plus nombreuses et mieux documentées. Les pouvoirs publics s'emparaient une seconde fois du sujet après le rapport FranceIA de mars 2017 avec le rapport de la Mission Villani de mars 2018. On n'a jamais autant évoqué l'IA dans les médias et même dans la fiction. Quelques nouveaux marchés verticaux ont été ajoutés. J'ai aussi réintégré une part du contenu de divers articles que j'ai pu publier entre octobre 2017 et avril 2018, sur la dimension créative de l'IA, sur l'IA émotionnelle, sur l'IA et la cybersécurité et sur l'IA symbolique.

La **quatrième édition** de novembre 2019 (624 pages) mettait à jour l'ensemble du document, avec quelques nouvelles parties (AIOps, art, usages de l'IA dans la recherche, les usages dans les achats, les fake news) et une sérieuse évolution des parties sur l'aviation, sur la distribution et sur l'éthique de l'IA. Cette mise à jour intervenait alors que les solutions d'IA devenaient pléthoriques, passant à l'ère de l'hyperchoix dans l'IA.

Cette **cinquième édition** de février 2021 (742 pages) grossit naturellement de manière homothétique dans toutes ses parties, avec une forte actualisation dans le deep learning, la vision et le traitement du langage, dans la partie matérielle, notamment dans la dimension énergétique de l'IA, dans les transports avec l'ajout de la mer, dans la santé et un peu partout ailleurs.

Le message clé à retenir est que l'intelligence artificielle n'est pas un produit. Elle ne se présente généralement pas sous la forme de logiciels packagés traditionnellement comme un traitement de texte, une application mobile ou un système d'exploitation. Il n'y a pas de logiciels d'intelligence artificielle mais des solutions logicielles et matérielles qui exploitent des briques d'intelligence artificielle variées qui s'appuient sur plusieurs dizaines de briques logicielles différentes qui vont de la captation des sens, notamment audio et visuel, à l'interprétation des informations, en passant par le traitement du langage et l'exploitation de grandes bases de données et de connaissances structurées ou non structurées. L'IA interagit avec ses utilisateurs via toutes les interfaces imaginables : par le texte (chatbots), les graphiques (reporting, dataviz), le tactile (dans les smartphones et tablettes), et même physiquement (avec la robotique). La création et l'intégration de solutions d'IA sont encore une affaire de bricolage et de tâtonnements, si ce n'est d'un véritable artisanat.

Derrière chaque prouesse d'une nouvelle IA se cachent aussi des Humains qui ont travaillé d'arrache-pied, ont développé et testé diverses techniques, échoué, réessayé, et d'autres qui ont souvent labellisé laborieusement des données d'entraînement. Il y a toujours de la sueur bien humaine derrière les nouvelles IA qui défraient la chronique, même pour celles qui sont dites créatives ! L'IA pure et non-humaine n'existe pas encore.

Nous en sommes toujours à l'âge de pierre, avec un peu plus d'une soixantaine d'années de recul sur la question et une vingtaine d'années pour ce qui est du deep learning. Les chercheurs du domaine continuent cependant de faire avancer la discipline. Le passage de la recherche à la production est de plus en plus rapide, les outils de développement de l'IA permettant de les mettre en pratique assez rapidement pour peu que les bons jeux de données soient disponibles. Or les jeux de données sont nombreux dans les entreprises ou en open data pour entraîner ses modèles ! D'autres essayent de faire en sorte que les IA nécessitent moins de données d'entraînement et veillent également à réduire son empreinte énergétique.

Dans mes différentes lectures scientifiques, en particulier autour du calcul quantique, j'ai découvert aussi la réalité des limites scientifiques et théoriques de l'informatique. Celle-ci ne peut pas résoudre tous les problèmes qui se présentent. Il existe des problèmes trop complexes pour les ordinateurs, même quantiques.

Des problèmes sont dits « indécidables » car ils n'ont pas de solution, ou tout du moins pas de solutions parfaites. L'IA peut aider à trouver des solutions optimales ou non et c'est au libre arbitre de l'Homme de les choisir. Cela permet de mieux comprendre pourquoi l'IA surhumaine qui pourrait tout prévoir, calculer et créer est un mythe. Je fais aussi la distinction entre les approches probabilistes de l'IA qui interprètent le monde par les données (approche top-bottom) et les outils mathématiques et informatiques qui permettent de simuler des phénomènes physiques à partir des équations qui les régissent (approche bottom-up). Les deux se complètent et évoluent en parallèle.

L'IA est un grand tonneau des Danaïdes scientifique. On n'arrive jamais à tout comprendre et à tout appréhender des techniques et domaines d'applications de l'IA. Chercher un "expert en IA"[2] revient maintenant à demander "un expert en logiciels" ou un "expert en informatique" sans compter le top avec "l'expert en transformation digitale".

Contrairement à un lieu commun répandu, les techniques et méthodes de l'IA évoluent sans cesse. Ce n'est pas qu'une question de puissance de machine ou de volume de données. Certains vieux de la vieille de l'IA considèrent qu'elle évolue très lentement et que rien n'a changé depuis leurs études. Ça me semble bien exagéré, même après avoir dé-marketinguisé les progrès récents de l'IA.

Selon certains, il faudrait un doctorat en IA pour pouvoir créer une solution intégrant de l'IA. C'est peut-être vrai aujourd'hui dans certains cas. Heureusement, de nombreux outils de développement et d'intégration arrivent sur le marché qui permettent à des développeurs moins qualifiés, voire même à des cadres, de créer eux-mêmes des solutions intégrant des briques d'IA, notamment dans le domaine du machine learning dédié à l'interprétation des données. C'est avec sa démocratisation que l'on peut évaluer l'évolution de la maturité d'une technologie émergente.

Je ne vais pas vous faire croire que j'ai tout compris à l'IA. Ce n'est pas le cas et des domaines de l'IA m'échappent encore. J'ai plein d'interrogations diverses, en particulier autour des techniques de traitement du langage et de leur dimensionnement, notamment sur les techniques les plus récentes (transformers, compositionnalité)[3] ! J'explore aussi les évolutions du champ de la représentation des connaissances et du raisonnement automatique, l'un des domaines les plus ardus de l'IA[4].

Cet ebook adopte un découpage en neuf grandes parties qui est adapté à la compréhension à la fois des techniques de l'IA, de ses outils de développement et, surtout, de ses usages dans les entreprises :

- **Histoire et sémantique de l'IA :** qu'est-ce que l'IA ? Qui a créé la discipline ? D'où vient cette appellation ? Pourquoi personne n'est d'accord sur le sens qu'il faut lui donner ? Comment l'IA est-elle segmentée d'un point de vue technique ? Quels sont ses grands courants intellectuels ? Comment cette discipline nouvelle a-t-elle progressé depuis les années 1950 ? Pourquoi a-t-elle connu deux grands hivers et qu'est-ce qui explique la dynamique actuelle ? Est-elle durable ? Où en est-on aujourd'hui ? Comment l'IA se compare-t-elle à l'intelligence humaine ?

[2] Voir Confession of a so-called AI expert de Chip Huyen, juillet 2017.

[3] Comme comprendre et expliquer dans le détail le fonctionnement des réseaux de neurones à mémoire de type LSTM.

[4] Voir Que devient l'IA symbolique ?, que j'ai publié en avril 2018.

- **Algorithmes et logiciels de l'IA** : quelles sont les principales briques mathématiques et algorithmiques de l'IA ? Le raisonnement automatique et les systèmes experts, et pourquoi en parle-t-on moins que pendant les années 1980 ? Quelles sont les techniques et applications du machine learning, des réseaux de neurones et du deep learning ? Les progrès récents viennent-ils du logiciel, du matériel ou des données ? Quels sont les outils de développement et de création d'applications de l'IA et pourquoi la majorité sont-ils open source ? Comment les briques d'intelligence artificielle progressent-elles ? Quid de l'intelligence artificielle générale ? Est-ce un fantasme ? Peut-on facilement reproduire le fonctionnement du cerveau humain ? Quels sont les projets allant dans ce sens et peuvent-ils aboutir ?
- **Données de l'IA** : quel est le rôle des données dans l'IA ? D'où viennent-elles ? Quelles sont les données ouvertes exploitables par l'IA ? Pourquoi les données peuvent-elles générer des biais dans les IA ? Comment les évite-t-on ? Quels sont les capteurs qui alimentent les données de l'IA ? Comment prépare-t-on et labellise-t-on les données du machine learning et du deep learning ? Comment les entreprises peuvent-elles valoriser leurs propres données ? Quelles sont les données ouvertes disponibles pour alimenter ses propres IA ? Comment audite-t-on les solutions de machine learning ?
- **Infrastructure de l'IA** : quelles sont les ressources matérielles qui font avancer l'IA ? Comment évolue la loi de Moore ? Quel est le rôle des GPU et des processeurs neuromorphiques dans l'IA ? Comment se distinguent-ils et comment les classifier ? Qui sont les acteurs de ce marché ? Pourquoi il y-a-t-il une grande différence entre l'entraînement d'une IA et son exécution dans la consommation de ressources matérielles ? Comment réduire l'empreinte énergétique de l'IA ? Est-ce que la photonique et l'informatique quantique auront un impact sur l'IA ? Comment sont gérées les ressources en cloud de l'IA ainsi que du côté des systèmes embarqués ? Comment architecturer les solutions d'IA ? Cette partie couvre aussi le champ des AIOps, celui de la RPA et enfin les questions de cybersécurité.
- **Applications génériques de l'IA** : quelles sont les applications génériques et horizontales de l'IA, dans le traitement de l'image, du langage, dans la robotique, dans le marketing, les ressources humaines, ainsi que la finance et les achats.
- **Applications métiers de l'IA** : quelles sont les grandes applications et études de cas de l'IA selon les marchés verticaux comme les transports, la santé, la finance, l'assurance, l'industrie, les utilities, le BTP et l'immobilier, la distribution, les médias, le tourisme, l'agriculture, l'éducation, le luxe, les services et le conseil, les métiers juridiques, les services publics, la défense et le renseignement. Pourquoi certains de ces marchés sont plus dynamiques que d'autres ? Comment les startups permettent aux entreprises d'innover dans ces différents marchés[5] ?
- **Acteurs de l'IA** : quelle est la stratégie et quelles sont les offres en IA des GAFAMI étendus, dont IBM, Google, Microsoft, Facebook, SalesForce, Oracle et plein d'autres encore ? Quid des Chinois ? Comment certains de ces acteurs se déploient-ils de manière verticale ? Comment se développent les startups en général et puis celles de l'écosystème français en particulier ? Comment évaluer la valeur ajoutée en IA des startups et autres acteurs de l'écosystème ? Comment les solutions d'IA sont-elles commercialisées ? Quelle est la part qui relève de produits et celle qui dépend des services et des données ?

[5] Je cite un très grand nombre de startups dans ce document. Il se peut que telle ou telle startup soit en déclin, ait fait un pivot ou n'existe plus. C'est la vie habituelle des startups. Je corrige le document au fil de l'eau lorsque nécessaire. Prévenez-moi !

- **IA et société** : les points de vue et études sur l'impact potentiel de l'IA sur l'emploi, les métiers et sur la société en général. Quelles sont les limites des prédictions ? Comment éviter de se faire robotiser ? Comment se préparer au niveau des compétences ? Quelles sont les grandes lignes de l'impact de l'IA sur la politique et les politiques de l'IA en France et ailleurs dans le monde ? Quel est l'état de la géopolitique de l'IA ? La Chine va-t-elle nous envahir avec son IA ? L'Europe est-elle condamnée, prise en sandwich entre les USA et la Chine ? Quelle est la situation de la France ?
- **IA et entreprise** : comment les entreprises peuvent-elles intégrer l'IA dans leur stratégie ? Quelles sont les bonnes méthodes et pratiques ? Comment benchmarker les solutions d'IA ? Comment s'organiser ? Comment gérer les compétences ? Quelles questions juridiques se posent dans l'adoption de l'IA ? Comment va évoluer le métier de développeur ? Comment se former en général ?

Ce document s'appuie toujours à la fois sur des rencontres avec des chercheurs, entrepreneurs et entreprises et surtout, sur une recherche bibliographique extensive. La littérature disponible sur le sujet est abondante, notamment les excellents cours de nombreuses universités. C'est la magie d'Internet quand on prend le temps de creuser ! L'abondance de la documentation scientifique permet notamment de faire le pont entre les effets d'annonces et la réalité sous-jacente.

Bonne lecture !

Olivier Ezratty, février 2021

Histoire et sémantique de l'IA

L'intelligence artificielle s'inscrit dans une Histoire humaine faite de machines qui démultiplient la puissance humaine, d'abord mécanique, puis intellectuelle avec de l'automatisation et une recherche effrénée de puissance et d'effets de levier. Elle s'alimente de nombreux mythes, de celui du golem qui mène aux robots ou Frankenstein et à celui d'Icare[6].

Tout cela alors qu'il n'existe même pas de consensus sur la définition de ce qu'est l'intelligence artificielle et même humaine[7] ! Elle est source d'une interminable bataille sémantique[8] qui touche notamment les startups du numérique[9]. Cela s'explique notamment par la difficulté que l'Homme a pour définir sa propre intelligence, sa conscience et ses mécanismes de raisonnement. Nous ne savons pas encore tout du fonctionnement du cerveau biologique. Mais il est vain de vouloir le copier à l'identique. Il doit servir de source d'inspiration pour faire différemment et mieux. Mais cela présente des limites : les avions ne battent pas de l'aile et n'ont pas de plume ! Les logiciels et l'IA ne sont donc pas condamnés à mimer le cerveau humain. Cela doit en être un appendice plutôt qu'une copie.

Cohabitent aussi des définitions étroites et larges de l'IA. Pour certains, seul le deep learning est digne de faire partie de l'IA et les moteurs de règles, pas du tout. Comme si seules les technologies un peu magiques dans leur apparence pouvaient faire partie de l'IA.

Pour nombre de spécialistes du secteur et votre serviteur, toutes les technologies de l'histoire de l'IA en font partie[10]. Certaines ont une dimension anthropomorphique comme la vision artificielle ou les agents conversationnels, d'autres, beaucoup moins lorsqu'elles analysent de gros volumes de données pour identifier des corrélations, des tendances ou faire des prévisions.

Des composantes isolées de l'intelligence humaine sont déjà intégrées dans les machines avec des capacités de calcul brutes et de mémoire qui dépassent celles de l'Homme depuis des décennies. Souvent, c'est la « force brute » qui permet aux IA de gagner à divers jeux de société. Le numérique a ainsi toujours utilisé la force brute pour dépasser l'homme, d'abord en capacités de calcul, puis de mémoire.

---

[6] Voir Tout ce que vous pensez savoir sur l'intelligence artificielle est complètement faux, où tout n'est pas forcément vrai, de Mathilde Rochefort, septembre 2016.

[7] Voir On Defining Artificial Intelligence par Pei Wang, 2019 (37 pages) qui propose une analyse sémantique fine de la notion d'intelligence artificielle. Il distingue plusieurs formes d'abstraction de l'intelligence artificielle : une abstraction structurelle (l'IA émule le cerveau humain), une abstraction comportementale (on s'en tient aux apparences, comme dans le test de Turing), une abstraction par les capacités (à résoudre des problèmes donnés), par les fonctions (en segmentant les fonctions cognitives humaines comme la perception, la planification ou le raisonnement) et par les principes (qui sous-tend une notion d'optimisation sous contraintes).

[8] Voir l'ouvrage « Intelligence artificielle – vers une domination programmée ? », de Jean-Gabriel Ganascia, seconde édition publiée 2017 d'une première édition datant de 1993, qui raconte très bien les débuts et le parcours de l'IA comme science. Pour Yann Le Cun, l'IA est est « la capacité, pour une machine, d'accomplir des taches généralement assumées par les animaux et les humains : percevoir, raisonner et agir. Elle est inséparable de la capacité à apprendre, telle qu'on l'observe chez les êtres vivants ».

[9] La querelle sémantique qui atteint l'univers des startups est toujours d'actualité. Celles-ci feraient de « l'IA washing », peignant aux couleurs de l'IA des solutions qui n'en contiennent pas forcément. Réflexion faite, cette notion d'IA washing est exagérée. Ce n'est pas parce que certaines utilisent des briques technologiques prêtes à l'emploi qu'elles ne font pas d'IA ou que leur solution n'intègre pas d'IA. C'est un peu comme si on disait qu'un site web réalisé en Wordpress avec un thème standard au lieu d'être développé avec son propre framework maison n'était pas "de l'Internet". Reste à définir "une IA", qui est toujours un assemblage de plusieurs composantes (données, algorithmes, hard, savoir faire métier) et à ausculter les startups en examinant le CV de leurs équipes techniques, leurs solutions et les données qui les alimentent. Ce qui permet de faire un premier tri.

[10] Voir cette intéressante histoire de l'IA : The Quest for Artificial Intelligence - A history of ideas and achievements de Nils Nilsson, Stanford University (707 pages). L'ouvrage date d'avant la grande vague du deep learning.

On peut opérer un jugement de Salomon en rappelant qu'artificiel est le contraire de naturel. Le débat sur la "vraie" IA est sans fin. Il n'y aura probablement jamais d'IA équivalente à l'intelligence humaine, ne serait-ce que parce que celle-ci est construite sur un substrat biologique qui l'alimente avec son cerveau, ses sens, son système hormonal, ses muscles, son squelette, ses besoins vitaux, une relation au temps et un processus de développement bien particulier. Le cerveau est lent mais massivement connecté et parallèle. Il se construit lentement par l'expérience dans le monde réel.

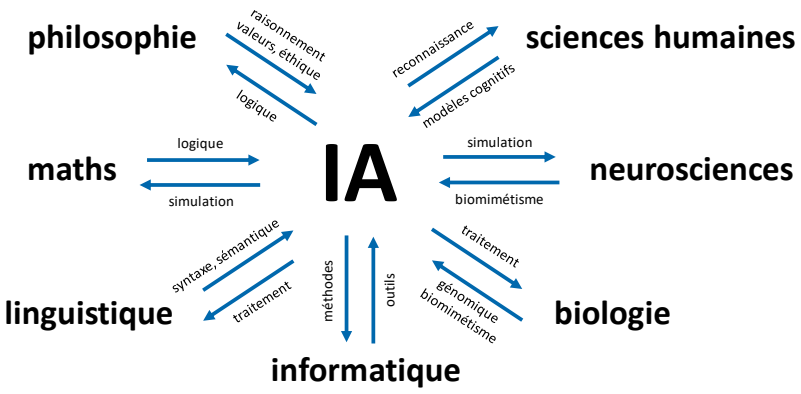
C'est en particulier le cas d'une partie méconnue du cerveau : le cervelet, qui comprend plus des deux tiers des neurones de notre cerveau alors qu'il n'en représente que 10% du poids. Celui-ci co-gère les mouvements appris avec le cortex. Quantitativement, une bonne part de notre intelligence est consacrée à la capacité à nous mouvoir dans l'espace. C'est un point commun que nous avons avec une bonne part des mammifères qui nous dépassent parfois de ce point de vue-là. Les robots ont d'ailleurs bien du mal à imiter cette compétence. La partie cognitive du cerveau ne représente qu'à peine 10% des neurones du cerveau. C'est surprenant [11]! Cela n'enlève rien à l'intelligence humaine qui est plus que diverse.

D'un point de vue sémantique, il faut peut-être s'échapper de la définition anthropomorphique de l'intelligence. Les modèles cognitifs humains ne sont pas forcément ceux que nous avons besoin de reproduire[12]. Les intelligences sont diverses. La notion d'intelligence intègre la capacité à comprendre, apprendre et à s'adapter à des situations nouvelles. On retrouve à des niveaux variables ces capacités dans des systèmes non vivants ou des systèmes vivants très simplifiés par rapport au cerveau humain. Les fourmis n'ont que 250 000 neurones en tout mais font preuve d'une prodigieuse intelligence collective, que l'on peut simuler dans l'IA avec des systèmes et réseaux multi-agents.

L'intelligence de l'Homme est très diverse, entre le citoyen moyen et le scientifique de génie et entre ma médiocre maîtrise de la flûte à bec et les meilleurs joueurs de violon du monde, ou Jimmy Page et Jimmy Hendrix à la guitare ! Rien n'empêche de créer d'autres formes d'intelligences que l'intelligence humaine, des alter-intelligences ! Les machines démultiplient la force physique de l'Homme. A charge pour l'IA de démultiplier son intelligence.

L'intelligence artificielle est aussi une discipline de l'informatique intimement liée à d'autres domaines scientifiques.

Avec les mathématiques, la logique et les statistiques qui lui servent de base théorique, les sciences humaines (sciences cognitives, psychologie, philosophie, linguistique, …) et les neurosciences qui aident à reproduire des composantes de l'intelligence humaine par biomimétisme[13], et enfin, les technologies matérielles qui servent de support physique à l'exécution des logiciels d'IA.



[11] C'est le paradoxe de Moravec selon lequel les facultés humaines les plus difficiles à reproduire sont celles qui sont en apparence les plus simples pour l'Homme, à savoir ses capacités motrices.

[12] Comme ce qu'évoque Why Cognitive Technology May Be A Better Term Than Artificial Intelligence par Kathleen Walch, 2019.

[13] Voir De quelles façons l'intelligence artificielle se sert-elle des neurosciences ? par Frédéric Alexandre, octobre 2019.

L'IA est un pan entier de l'informatique avec sa diversité, ses briques technologiques, ses méthodes, ses assemblages et solutions en tout genre. Elle est maintenant imbriquée dans quasiment tous les pans du numérique.

Dans le grand public, des briques d'IA sont déjà utilisées au quotidien sans que les utilisateurs le remarquent. C'est par exemple le cas des systèmes de détection et de suivi des visages dans la mise au point des photos et vidéos des smartphones. Et puis bien entendu, avec les agents conversationnels comme Google Assistant, Apple Siri et Amazon Alexa.

Hauts et bas de l'IA

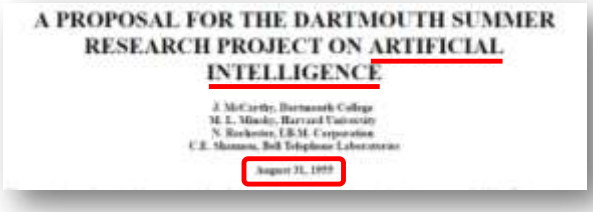
L'IA puise ses sources dans le concept théorique de *calculus ratiocinator* de **Gottfried Wilhelm Leibnitz** (circa 1671), les travaux de **George Boole** en 1854 sur la logique symbolique, la machine et le fameux test d'**Alan Turing** que l'on ne présente plus (1935 et 1950), les neurones formels de **Warren McCulloch** et **Walter Pitts** (1943), l'architecture de **John Von Neumann** qui sert encore de base à celle des ordinateurs traditionnels d'aujourd'hui avec unité de contrôle, unité de calcul, mémoire pour les programes et les données (1945) ou encore le théorème de l'information de **Claude Shannon** (1949).

L'histoire moderne de l'intelligence artificielle a cependant véritablement pris son envol lors du **Summer Camp de Darmouth**, organisé entre le 18 juin et le 17 août 1956 à Hanover dans le New Hampshire aux USA. Le groupe de travail fondateur du summer camp de Darmouth comprenait **John McCarthy** (29 ans), **Claude Shannon** (40 ans) et **Oliver Selfridge** (30 ans) tous trois du MIT, **Marvin Minsky** (29 ans) de Princeton, **Allan Newell** (29 ans) et **Herbert Simon** (40 ans) de Carnegie Tech, puis **Arthur Samuel** (55 ans) et **Nathaniel Rochester** (37 ans), tous deux d'IBM. Ce dernier avait conçu l'IBM 701, le premier ordinateur scientifique d'IBM, à lampes, lancé en 1952 et commercialisé à 19 exemplaires. On remarquera dans ces lignes que l'histoire de l'IA évolue parallèlement et simultanément à celle de l'informatique classique à partir des années 1950.

Summer Camp de Darmouth

USA New Hampshire – été 1956

définit le périmètre d'investigation de l'IA

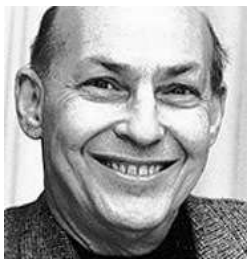

A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

« The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. »

- Automatic computers
- Use a language
- Neuron nets
- Theory of the size of calculation
- Self-improvements
- Abstractions
- Randomness and creativity



Marvin Minksy, MIT
1927-2016

Claude Shannon, Bell
1916-2001

John McCarthy, MIT
1927-2011

Nathaniel Rochester, IBM
1919-2001

Le Congrès de Darmouth était une sorte de hackathon intellectuel de huit semaines. L'objectif était de réfléchir aux concepts permettant de reproduire dans des machines diverses composantes de l'intelligence humaine de base comme la maîtrise du langage, la vision et le raisonnement. Environ une trentaine de personnes y ont participé en tout. L'Histoire a donné une importance surdimensionnée à cet événement qui était plutôt un point de départ, une étape, plus qu'un accomplissement.

Voyage éternel ou aboutissement ?

L'expression « intelligence artificielle » fut couchée sur papier le 31 août 1955 par **John McCarthy**[14] dans la note de 13 pages proposant l'organisation du Summer Camp de Darmouth[15]. Elle recouvre les théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence. Il s'agit de sciences et technologies qui permettent d'imiter, d'étendre et/ou d'augmenter l'intelligence humaine avec des machines. Aucune IA n'a à ce jour les capacités intégratives de l'intelligence humaine. A noter un biais humain courant d'interprétation de la prouesse des machines : si on en comprend la méthode, ce n'est plus de la magie et ce n'est plus intelligent. L'intelligence est souvent mieux appréciée si elle est inexplicable !

Dans la note d'août 1955, l'intelligence artificielle était présentée comme une conjecture. Quelque chose qui n'est pas démontré. C'est toujours le cas à ce jour[16] ! A cette date, l'informatique est alors un marché naissant avec à peine quelques centaines d'ordinateurs dans le monde, fonctionnant encore avec des lampes. 1955 était l'année de l'apparition des premiers mainframes à transistors avec le TRADIC des Bell Labs et l'IBM 702 dont seulement 14 exemplaires ont été fabriqués. C'est l'IBM S/360 qui va véritablement faire émerger le marché des mainframes. Sa commercialisation démarrera en 1965 ! Il sera notamment utilisé à Houston dans la salle de contrôle du programme Apollo. Le PDG d'IBM de cette époque est Thomas Watson Jr, le fils du fondateur d'IBM Thomas Watson. Fondation en 1924 à partir de la société C.T.R. qui avait consolidé l'activité de quatre sociétés dont la Tabulating Machine Company créée en 1896 par Herman Hollerith, l'inventeur de la mécanographie à base de cartes perforées.

Selon Grace Solomonov, la femme de **Ray Solomonov**, l'un des participants du Summer Camp de Darmouth, le terme d'IA a été choisi par John McCarthy parce qu'il était neutre par rapport aux sciences existantes comme la cybernétique, issue en 1948 des travaux du mathématicien et philosophe **Norbert Wiener** (1894-1964). La science-fiction de l'époque inspira aussi son mari[17].

Contrairement à nombre de nouvelles expressions du numérique (client serveur, big data, objets connectés, Internet des objets, blockchain) qui trouvent leur origine dans le marketing de sociétés privées, l'IA est donc une appellation créée par un chercheur ! Elle lui permettait d'installer son domaine dans le paysage et d'éviter son assimilation à des disciplines voisines comme les mathématiques, les statistiques ou l'informatique. C'était donc une forme de déclaration d'indépendance d'une nouvelle discipline scientifique.

Plus de 60 ans plus tard, l'IA décrit aussi bien le champ du possible d'aujourd'hui dans ces domaines que la quête permanente et insatisfaite de l'incorporation des différents aspects de l'intelligence humaine dans des machines. Si l'IA a atteint temporellement l'âge de la retraite, elle reste encore adolescente et brouillonne en pratique.

---

[14] Pour la petite histoire, 1955 est aussi l'année de la naissance de Steve Jobs et Bill Gates.Tout un symbole ! Dans la fiction, c'est aussi l'année du retour dans le passé de Marty McFly dans Back to the Future (1985).

[15] Le document historique est A proposal for the Darmouth summer research project of artificial intelligence, 31 août, 1955. Il prévoyait un budget de $13500, les deux tiers correspondant aux salaires des participants, situés entre $600 et $700 par mois, ce qui équivaudrait à environ $6000 d'aujourd'hui, charges comprises. La note a été envoyée à la Fondation Rockefeller le 2 septembre 1955. Le bilan de Solomonov n'était pas enthousiasmant : "*The research project wasn't very suggestive. The main things of value : 1) wrote and got report reproduced (very important) 2) Met some interesting people in this field 3) Got idea of how poor most thought in this field is 4) Some ideas: a) Search problem may be important. b) These guys may eventually invent a Turing Machine simply by working more and more interesting special problems*".

[16] Précisément : "*The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. […] Topics to study: automatic computers and programs for them; programming computers to use a language; neuron nets; machine self-improvement; classifying abstractions; and two areas particularly relevant to Solomonoff's work, though they would not be part of AI for many years — how to measure the complexity of calculations, and randomness and creativity*".

[17] Elle a rédigé en 2006 un document qui relate l'Histoire détaillée de ce Summer Camp : Ray Solomonoff and the Dartmouth Summer Research Project in Artificial Intelligence, 1956 (28 pages).

L'appellation la plus appropriée serait peut-être celle d'**intelligence humaine augmentée**[18], l'IA étant principalement destinée à démultiplier les capacités de l'Homme, comme tous les outils numériques jusqu'à présent, même si dans certains cas, l'IA peut effectivement se substituer aux travaux de l'Homme pour quelques tâches élémentaires comme, à relativement moyen terme, la conduite de véhicules[19]. Cela se rapproche de la notion d'exo-darwinisme promue par le philosophe **Michel Serres** (1930-2019)[20]. Dans le raisonnement automatisé, l'IA est censée apporter une rationnalité dont l'Homme ne fait pas toujours preuve. Là encore, nous sommes dans l'ordre de la complémentarité.

Autre élément de compréhension : se rappeller que l'IA imite et simule l'intelligence humaine, et des mammifères et surtout pour ce qui est de la partie non cognitive (sens, capacité de mouvement). Elle sert aussi d'outil d'analyse et de compréhension de l'intelligence humaine.

Il est fascinant de confronter deux avis en apparence complètement opposés : celui de **Luc Julia** selon lequel l'IA n'existe pas, au sens, d'une IA anthropomorphique imitant l'intelligence humaine, et celui de **Laurent Alexandre** selon lequel l'IA produirait des gilets jaunes, sachant qu'il faut chez lui souvent traduire IA par ruptures technologiques en général[21].

L'intelligence artificielle incarne finalement la conquête d'un **Graal** distant, ayant été à l'origine, sur son chemin, d'un tas d'avancées technologiques relativement distinctes et plutôt complémentaires de l'intelligence humaine[22]. Celle-ci est encore unique dans la capacité à réagir avec discernement face à des situations nouvelles, à tirer profit de circonstances fortuites, à discerner le sens de messages ambigus ou contradictoires, à juger de l'importance relative de différents éléments d'une situation, à trouver des similitudes entre des situations malgré leurs différences, à établir des distinctions entre des situations malgré leurs similitudes, à synthétiser de nouveaux concepts malgré leurs différences ou à trouver de nouvelles idées[23].

---

[18] Voir Why AI Should Rightfully Mean Augmented Intelligence, Not Artificial Intelligence par Joe McKendrick, juin 2019.

[19] Voir Intelligence artificielle : en finir avec les mirages de Vincent Champain, avril 2018 qui évoque cette notion d'intelligence augmentée.

[20] Voir Exo-Darwinisme Numérique, Olivier Ezratty, octobre 2007, qui pointe sur une conférence de Michel Serres où il expose brillamment ce concept.

[21] L'illustration de droite est une interview Pour moi l'IA n'existe pas de Luc Julia, parue dans CBNews en juin 2018. Luc Julia en a fait un livre publié début 2019. Voir son interview « L'intelligence artificielle n'existe pas » : interview de Luc Julia, le cocréateur de Siri, par Anne Cagan, janvier 2019.

[22] On pourrait dire qu'il en va de même des cancérologues dont le métier est de guérir le cancer et qui n'y arrivent pas forcément.

[23] La source de cette énumération est le cours d'intelligence artificielle d'Olivier Boisard.

Faut-il pour autant enterrer l'IA d'aujourd'hui comme le font certains qui indiquent qu'elle est dans l'impasse, qu'elle représente un grand mensonge si ce n'est une escroquerie[24] ? Comme nous leverrons dans cet ouvrage, l'IA d'aujourd'hui rend bien des services, sous certaines conditions. Si elle n'est pas anthropomorphique, ce n'est pas bien grave tant qu'elle peut compléter l'intelligence humaine et dans tous les secteurs d'activité.

Bases conceptuelles

Deux ans après le summer camp de Darmouth avait lieu le **Congrès de Middlesex** (1958) au Royaume-Uni avec des contributions des principaux participants du Congrès de Darmouth, **Marvin Minsky** et **John MacCarthy** ainsi qu'**Oliver Selfridge**, lui aussi présent à Darmouth. L'objet des publications associées était la modélisation et la mécanisation des mécanismes de la pensée en particulier avec des logiques heuristiques. C'était de l'IA symbolique.

S'en suivirent des publications clés comme celles de **Marvin Minsky** qui jetait les bases théoriques de la programmation heuristique[25]. La même année, **Oliver Selfridge**, posait les bases des réseaux de neurones pour la reconnaissance des formes[26] (IA connexionniste) et **John McCarthy**, celle des systèmes experts[27] (IA symbolique). Ce dernier est aussi connu pour la création en 1958 du langage de programmation **LISP**, à la base de systèmes à base de logique formelle et de règles.

Les années 1960 furent une période active de recherche fondamentale, notamment au **MIT AI Lab**. Ces recherches étaient principalement financées par l'ARPA, l'agence de recherche du Pentagone créée en 1958, juste après l'épisode du lancement du satellite Spoutnik par l'URSS en 1957, devenue la **DARPA** en 1972, dotée d'un budget annuel de plus de $3B[28].

La recherche sur l'IA était surtout financée par les deniers publics. Encore aujourd'hui, une très grande partie des recherches les plus avancées sur l'IA aux USA sont financées par l'omniprésente DARPA et réalisées par des laboratoires de recherche d'Université ainsi que chez les GAFAMI. Via l'agence IARPA, les agences du renseignement comme la CIA et la NSA financent aussi la recherche en IA. Cela peut alimenter au passage les craintes sur les applications futures de l'IA, notamment au moment hypothétique où elle atteindrait le stade de l'AGI (IA généraliste).

En **France**, on peut noter quelques dates clés avec la création d'**Inria** en 1967, à l'époque IRIA[29], dans le cadre du plan calcul visant à apporter une indépendance à la France dans les calculateurs nécessaires au développement de la dissuasion nucléaire nationale.

---

[24] Voir L'intelligence artificielle est « dans l'impasse » par Vincent Bérenger, décembre 2019, « The wall » : les limites à venir de l'IA par Etienne Grass, février 2020, « L'intelligence artificielle est bien aujourd'hui une escroquerie ! » par Robert Bentz, novembre 2019, "AI is a lie" – O'Reilly par Jenn Webb, 2019 et How to recognize AI snake oil par Arvind Narayanan (21 slides).

[25] Voir Some Methods of Artificial Intelligence and Heuristic Programming de Marvin Minsky, 1958 (34 pages) qui fut approfondie peu après dans Steps Toward Artificial Intelligence de Marvin Minksy, 1960 (23 pages).

[26] Dans Pandemonium : a paradigm for learning, Olivier Selfridge, 1958 (21 pages).

[27] Voir Programming with common sense, John McCarthy. 1959 (15 pages).

[28] J'utilise la nomenclature US pour les montants en dollars. $3B veut dire trois milliards de dollars. En Euros, j'utilise 3Md€. Lorsque je cite une startup, j'indique généralement entre parenthèses son année de création, le pays d'origine et la totalité des financements obtenus, informations récupérées le plus souvent sur Crunchbase.

[29] Le IA signifiait Informatique et Automatisme, pas Intelligence Artificielle !

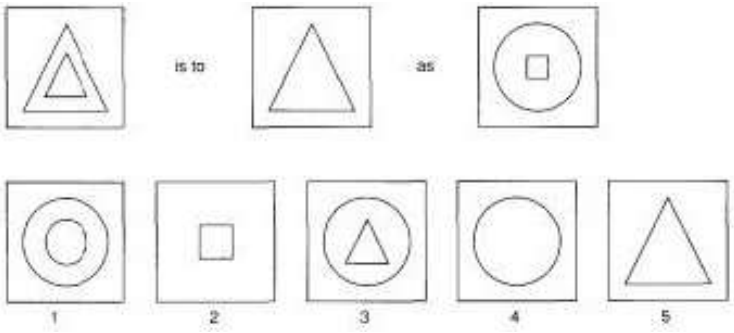
Le plan calcul avait aussi amené par Mecano industriel à la création en 1966 de la **CII** (Compagnie Informatique Internationale) consolidant l'activité de plusieurs PME du secteur. La CII fusionnait avec Honeywell Bull en 1975, devenu Bull en 1982, enfin absorbé par Atos en 2014.

La création du langage **Prolog** par Alain Colmerauer[30] et Philippe Roussel à l'Université d'Aix-Marseille de Luminy date de 1972. La recherche française en IA ne semble cependant avoir véritablement décollé qu'à la fin des années 1970. Elle s'est plutôt spécialisée dans l'IA symbolique, un domaine dont on entend peu parler depuis le tsunami mondial du deep learning lancé à partir de 2006. Il recouvre diverses techniques de modélisation des connaissances et du raisonnement.

Démonstrations de théorèmes

Les premiers travaux autour de l'IA symbolique portèrent sur la démonstration automatique de théorèmes, notamment en géométrie de base.

Le premier en date est le **Geometry Theorem Prover** d'Herbert Gelernter de 1959, un logiciel de démonstration de théorèmes de géométrie fonctionnant en chaînage arrière - de la solution jusqu'au problème - sur un IBM 704 à lampes et à partir d'une base de 1000 règles. Cela relevait d'une combinatoire assez rudimentaire. C'était plutôt prometteur.



Suivirent le **General Problem Solver** d'Allen Newell et Herbert Simon en 1959, l'**Integration Problems Solver** de James Slagles en 1963, le **Geometric Analogy Problems** de Tom Evans en 1968, qui traitait les problèmes de géométrie qui sont intégrés dans les tests de quotient intellectuel (*ci-dessus*) et puis l'**Algebra Problems Solver** de Daniel Bobrow en 1967. Tout cela bien avant les débuts de la micro-informatique ! Les méthodes créées servirent plus tard de base aux techniques de moteurs de règles et de systèmes experts qui connurent leur heure de gloire pendant les années 1980.

Premiers chatbots

On vit apparaître les ancêtres de catégories de solutions d'IA courantes aujourd'hui avec l'un des premiers chatbots, simulant un dialogue avec un psy, **ELIZA** entre 1964 et 1966. Il aurait inspiré Stanley Kubrick et Richard C. Clarke dans la préparation du scénario de « 2001 Odyssée de l'Espace » pour le personnage-machine HAL[31]. Après ELIZA, suivra le système **SHRDLU**, de Terry Winograd du MIT, l'un des premiers à comprendre le langage naturel en 1968 (vidéo).

[30] Alain Colmerauer est décédé en mai 2017.

[31] Voir The Real Technology Behind 2001's HAL par Paul Ceruzzi, mai 2018. L'écriture du scénario du film a démarré en 1964.

Ces premiers chatbots tenaient le coup pendant des conversations avec quelques échanges mais ne passaient pas le fameux **test de Turing**[32]. Malgré tout, ils n'ont pas à rougir vis-à-vis de nombreux chatbots contemporains, sauf peut-être pour les plus récents développés avec le modèle GPT-3 provenant d'OpenAI.

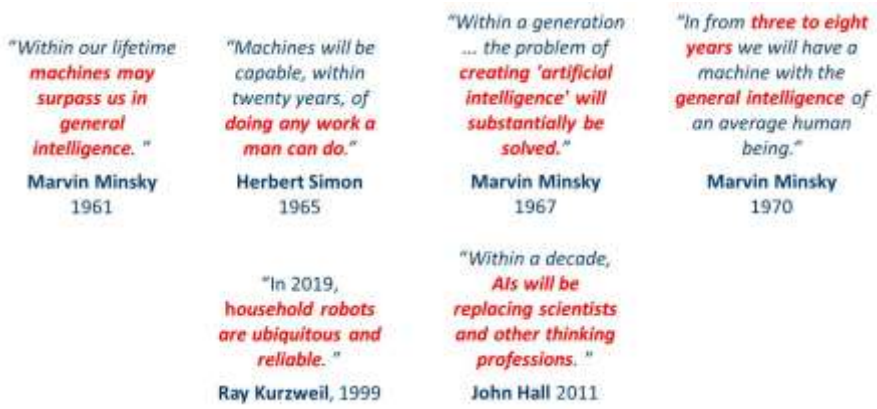
Surpromesses et premier hiver de l'IA

L'IA connut son premier "hiver" avec une réduction d'une bonne part de ses budgets de recherche à partir du début des années 1970, tant au Royaume-Uni qu'aux USA. En cause, les surpromesses des scientifiques du domaine et les lentes avancées associées.

Aux USA, le Congrès votait les **Mansfield Amendments** dans le Defense Procurement Act de 1969 puis en 1973, du nom d'un sénateur démocrate, demandant à ce que la recherche financée par l'ARPA, l'ancêtre de la DARPA, ait des applications directes dans l'armée. En gros, cela coupait les financements de la recherche fondamentale civile, dont celle de l'IA et en particulier les travaux de **BBN Technologies**, un sous-traitant de l'ARPA situé à Cambridge dans le Massachusetts, créé par des anciens chercheurs du MIT et filiale de Raytheon depuis 2009. On doit à BBN la mise en œuvre du réseau ARPANET et l'invention de l'email en 1971 ! Ironie de l'Histoire, par un phénomène de vases communicants, ces coupes budgétaires sur l'IA auraient contribué indirectement à la naissance de l'industrie de la micro-informatique au milieu des années 1970[33] !

En effet, alors que ce premier hiver de l'IA a duré jusqu'en 1980, nous avons eu pendant ce trou d'air la création du **Micral** en France (1973), de **Microsoft** (1975), d'**Apple** (1977), d'**Oracle** (1977), le Alto (1973) puis le Star de **Xerox** (1980) qui a ensuite inspiré Apple pour Lisa (1983) et le Macintosh (1984), puis enfin les préparatifs du lancement de l'**IBM PC** (1980-1981).

Au Royaume-Uni, cet hiver était la conséquence de la publication du **Rapport Lighthill** qui était destiné au **Science Research Council**, l'équivalent de notre Agence Nationale de la Recherche d'aujourd'hui. Il remettait en cause le bien fondé des recherches de l'époque en robotique et en traitement du langage[34].



C'est une approche bien curieuse quand on sait que les technologies informatiques matérielles sous-jacentes n'étaient pas encore bien développées à cette époque[35]. C'est un bel exemple de manque de vision long terme des auteurs de ce genre de rapport. En France, l'équivalent serait le fameux Rapport de Gérard Théry de 1994 sur les autoroutes de l'information qui dénigrait Internet et le Web.

Le rapport Lighthill mettait en avant les promesses trop optimistes des experts du secteur. Comme souvent, les prévisions peuvent être justes sur tout ou partie du fond mais à côté de la plaque sur leur timing comme le montrent les citations *ci-dessus*[36].

---

[32] Le test de Turing est décrit implicitement dans Computing Machinery and Intelligence, Alan Turing, 1950 (20 pages).

[33] Comme relaté dans la fiche Wikipedia sur la DARPA : https://en.wikipedia.org/wiki/DARPA.

[34] Voir Lighthill Report: Artificial Intelligence: a paper symposium, 1973.

[35] 1973 est l'année de l'apparition du premier micro-ordinateur de l'histoire, le français Micral de François Gernel et André Truong.

[36] Voir Machine who think de Pamela McCorduck, 2004 (584 pages) qui fait un inventaire intéressant des erreurs de prévisions en relatant les cinquantes premières années de l'Histoire de l'intelligence artificielle. Cet ouvrage librement téléchargeable est par ailleurs très instructif sur l'histoire de l'IA. Voir aussi cette timeline de History of artificial intelligence.

Herbert Simon et **Allen Newell** prévoyaient en 1958 qu'en dix ans, un ordinateur deviendrait champion du monde d'échecs et qu'un autre serait capable de prouver un nouveau et important théorème mathématique. Trente ans d'erreur de timing pour la première prévision et autant pour la seconde sachant qu'elle est toujours largement en devenir pour être générique !

Herbert Simon (*ci-contre*) prévoyait– toujours en 1958 – qu'en 1978, les machines seraient capables de réaliser toutes les activités intellectuelles humaines. Et la loi de Moore n'existait pas encore puisqu'elle a été énoncée bien après cette prévision, en 1965 et observée entre les années 1970 et 2010. En 1967, **Marvin Minsky** pensait qu'en une génération, tous les problèmes liés à l'IA seraient résolus. Deux générations plus tard, on en discute encore. Il prévoyait aussi qu'au milieu des années 1970, les ordinateurs auraient l'intelligence d'un homme moyen.

Reste à savoir ce qu'est un homme moyen. Et combien de robots peuvent courir un marathon ?

Les retards étaient manifestes dans la traduction automatique et dans la reconnaissance de la parole. Notons qu'Herbert Simon a été récompensé en 1978 par le Prix Nobel d'économie, pour ses travaux sur les rationalités de la prise de décision, après avoir gagné la fameuse médaille de Turing en 1975. Il n'existe pas encore de prix Nobel de la prévision ! Il faudrait d'ailleurs plutôt les attribuer à des personnes déjà décédées pour valider leurs prévisions au long cours !

Ces prévisions trop ambitieuses ont proliféré. Leurs versions actualisées tournent autour de la singularité et du transhumanisme : l'ordinateur plus intelligent que l'homme entre 2030 ou 2045 et l'immortalité ou une vie de 1000 ans pour les enfants qui viennent de naître !

Première renaissance avec les systèmes experts et nouvel hiver

Ce premier hiver a été suivi d'une période d'enthousiasme au début des années 1980 alimentée notamment par la vague des systèmes experts. Cet enthousiasme a duré moins d'une décennie. Une nouvelle vague de désillusions s'en est suivie autour des années 1990.

Notamment du fait de l'essoufflement de la vague des systèmes experts et l'effondrement associé du marché des ordinateurs dédiés au LISP[37], notamment ceux des startups **Symbolics** et **Lisp Machines**, toutes deux issues du MIT, la première ayant été lancée en 1980 et fermée en 1996. Ces deux startups se disputaient sur des questions de propriété intellectuelle. C'est ce qui a conduit Richard Stallman, lui aussi du MIT, à créer la **Free Software Foundation** en 1985 et à lancer le mouvement du logiciel open source. L'Histoire a des ramifications inattendues !

[37] Voir cet historique : Lisp machine dans Wikipedia.

L'autre raison du déclin de l'IA était que le matériel n'arrivait pas à suivre les besoins de l'IA, notamment pour traiter deux besoins clés : la reconnaissance de la parole et celle des images, très gourmandes en puissance de calcul, même avec les algorithmes de l'époque qui n'exploitaient pas encore les réseaux de neurones d'aujourd'hui. Il s'agissait des débuts de l'approche connexionniste (1980-1995).

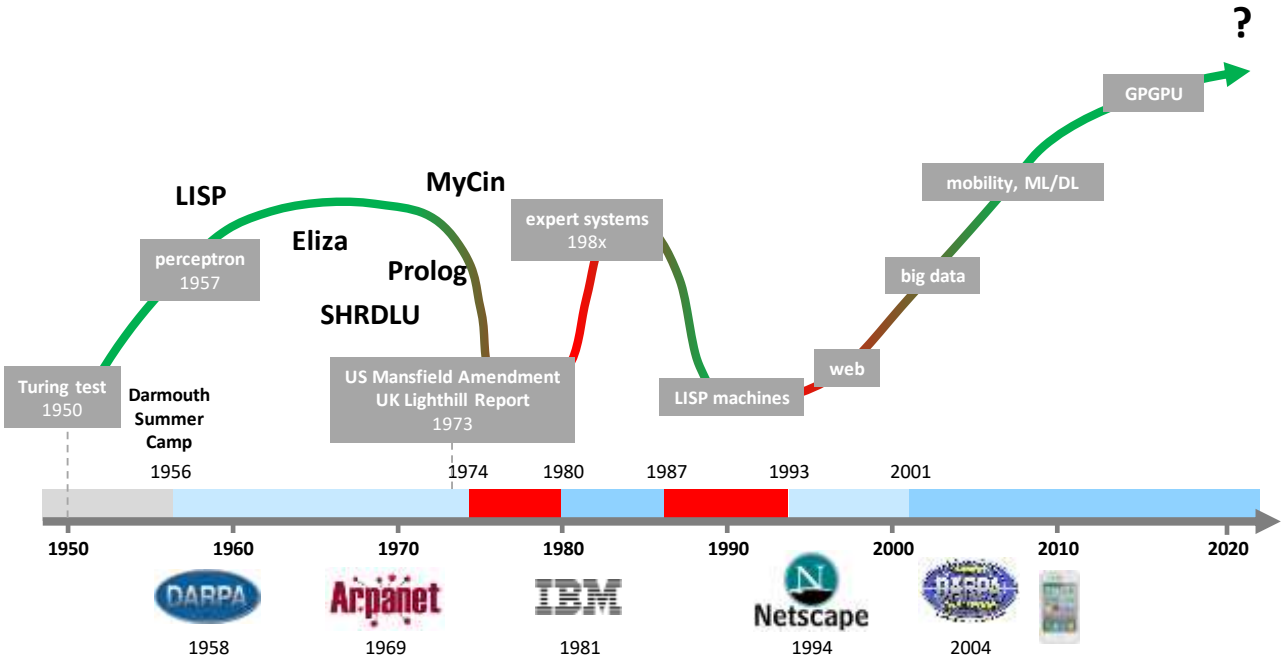
Lors des années 1980 avaient été lancés divers *gosplans* d'ordinateurs "*de cinquième génération*" dédiés aux applications de l'IA. Cela a commencé avec celui du **MITI Japonais,** lancé en 1981 avec un budget de $1B (un milliard de dollars), puis avec le projet anglais **Alvey** doté de £350M et enfin, avec le **Strategic Computing Initiative** lancée par la DARPA en 1983. Tous ces projets ont capoté et ont été clôturés discrètement.

Le projet du MITI visait à faire avancer l'état de l'art côté matériel et logiciel. Les japonais cherchaient à traiter le langage naturel, à démontrer des théorèmes et même à gagner au jeu de Go. Le projet a probablement pâti d'une organisation trop traditionnelle, hiérarchique et centralisée.

Pendant les années 1990 et 2000 ont émergé de nombreux projets de **HPC** (high-performance computers), assez éloignés de l'IA et focalisés sur la puissance brute et les calculs de simulation par éléments finis.

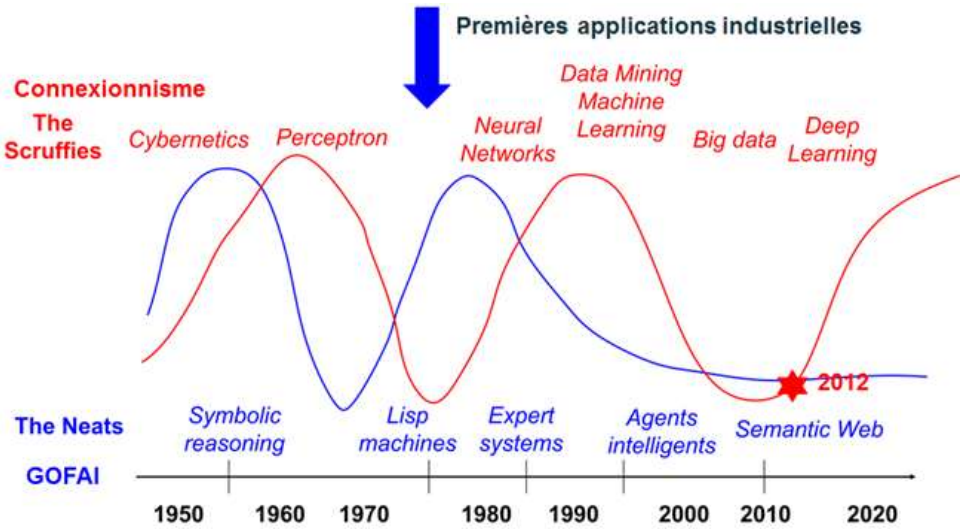
Ils étaient et sont encore utilisés pour de la simulation, notamment d'armes nucléaires, d'écoulements d'air sur les ailes d'avion ou pour faire des prévisions météorologiques. Les HPC de **Cray Computers** avaient été créés pour cela ! Cette société est restée indépendante jusqu'à son acquisition par SGI en 1996, ce dernier étant à son tour absorbé par HPE en juillet 2019.

Le marché des supercalculateurs a été renouvelé à partir de 2017 avec l'adoption des GPGPU de Nvidia. Ils sont ainsi devenus des machines aussi adaptées aux applications les plus exigeantes de l'IA.



Le schéma *ci-dessus* de mon cru illustre ces hauts et ces bas par décennies. En voici, *ci-dessous* une variante, provenant de Françoise Soulié-Fogelman et qui présente le décalage entre les étés et les hivers des IA symbolique et connexionniste.

Les réseaux de neurones et le deep learning ont connu un premier essor au milieu des années 1990 (évoqué dans cette partie sur les débuts du deep learning) avant de s'essouffler dans les années 2000 puis de redécoller véritablement à partir de 2012 avec l'arrivée de GPU de Nvidia.



Dernière renaissance de l'IA

Depuis le début des années 2000, et surtout depuis 2012, l'IA a été relancée grâce à diverses évolutions majeures :

Tout d'abord, les **progrès théoriques et pratiques** constants dans le machine learning, les réseaux de neurones et le deep learning. Nous aurons l'occasion de tous les évoquer plus en détail dans la seconde partie dédiée aux algorithmes et logiciels de l'IA.

L'augmentation de la **puissance du matériel** a ensuite permis de diversifier la mise en œuvre de nombreuses méthodes jusqu'alors inaccessibles. Et en particulier, l'usage du machine learning pouvant exploiter la puissance des machines autant côté calcul que stockage et puis, plus récemment, les réseaux neuronaux et le deep learning. Cette augmentation de puissance se poursuit inexorablement, malgré les limites actuelles de l'intégration des transistors dans les circuits intégrés. Une partie de ce document est aussi dédiée à cet aspect.

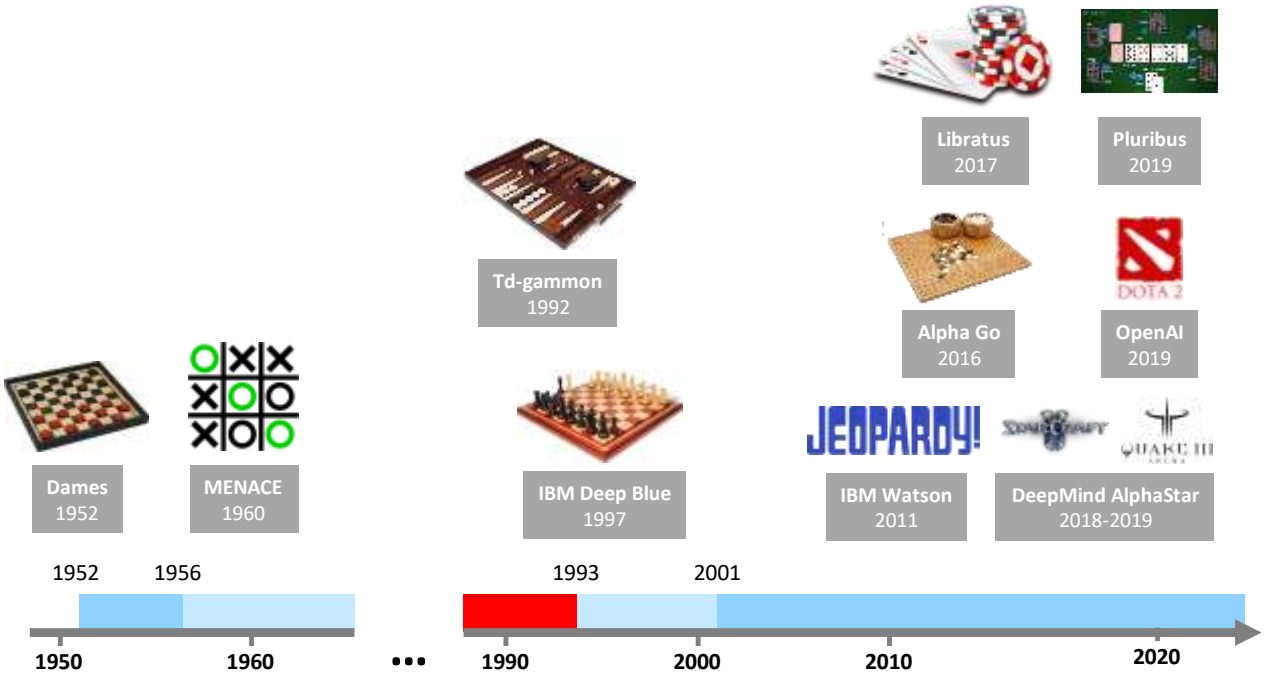
Les avancées de l'IA ont aussi été marquées par l'atteinte de diverses **étapes marquantes** comme la victoire d'IBM Deep Blue contre Kasparov en 1997 (après la victoire de ce dernier contre Deep Blue dans la moitié des parties jouées entre 1996 et 1997) puis d'IBM Watson dans Jeopardy en 2011. Enfin, début 2016, la victoire de Google DeepMind AlphaGo au jeu de Go contre le champion du monde. Les premiers jeux de société gagnés via l'IA étaient le jeu de dames (Checkers) et le tic-tac-toe dans les années 1950-1960. Il y avait eu près de 30 ans de calme plat dans le domaine des jeux de société.

Depuis, trois IA ont aussi gagné au jeu de poker[38], Libratus et DeepStack en 2017 puis Pluribus en 2019, face à plusieurs champions[39] ! Par rapport aux échecs ou au jeu de Go où le jeu est entièrement visible, la performance de ces IA tient d'une part au fait qu'elles agissent dans un environnement d'information incomplet et, d'autre part, au fait qu'elles peuvent moduler l'agressivité du jeu.

[38] Voir Artificial intelligence goes deep to beat humans at poker, mars 2017. La description technique de DeepStack, créé par des chercheurs canadiens et tchèques, est dans DeepStack: Expert-Level Artificial Intelligence in No-Limit Poker. Celle de Libratus, créé par Tuomas Sandholm et NoamBrown, de l'Université Carnegie Mellon de Pittsburgh est dans Libratus: The Superhuman AI for No-Limit Poker et pour la vulgarisation, dans cet article de Wired, Inside Libratus : the poker AI that out-bluffed the best humans. Dans les deux cas, il s'agissait de parties 1 contre 1. DeepStack et Libratus sont bien décrits dans cette présentation technique : Poker AI: Equilibrium, Online Resolving, Deep Learning and Reinforcement Learning de Nikolai Yakovenko (Nvidia), avril 2017. La prochaine étape sera d'intégrer à ces IA des capteurs sensoriels permettant de détecter les émotions des joueurs humains. A distance et avec une caméra, on peut détecter de fines variations dans les expressions et même la variation du pouls !

[39] Voir Facebook AI forces poker pros to fold in Texas Hold'em tourney par Qweenie Wong, 2019. Pluribus a été développé par Facebook et des équipes de l'Université Carnegie Mellon.

Fin 2017, le logiciel AlphaGo de DeepMind était adapté pour jouer et gagner au jeu d'échecs et à sa variante japonaise du Shogi en apprenant par lui-même à jouer[40]. En mai 2018, le laboratoire de recherche en IA de Facbebook (FAIR) publiait OpenGo, sa version open source d'une IA gagnant aussi au jeu de Go face à d'autres IA publiques, mais pas face à AlphaGo Zero qui n'est pas en open source[41].



DeepMind a aussi créé AlphaStar, une IA qui joue dans les jeux vidéo **Starcraft** et **Quake III** et battait en janvier 2019 les meilleurs joueurs du monde[42]. Et en octobre 2019, DeepMind faisait atteindre le plus haut niveau de jeu possible, le Grandmaster, pour son IA multi-agents pour **StarCraft II**[43].

OpenAI Five battait une équipe de joueurs humains dans Dota 2 en avril 2019 après avoir battu un seul joueur en 2017[44]. En 2020, ce même OpenAI créait un algorithm capable d'apprentissage pour résoudre un Rubik's Cube avec une main robotisée[45].

Tous ces succès récents s'appuient sur de l'IA connexionniste et du deep learning, de plus en plus associée à des réseaux d'agents entraînés par renforcement[46]. Des algorithmes génétiques ont été utilisés pour gagner à différents jeux Atari[47].

---

[40] Voir Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, décembre 2017 (19 pages) et DeepMind Achieves Holy Grail: An AI That Can Master Games Like Chess and Go Without Human Help, Philipp Ross, décembre 2018.

[41] Voir Facebook Open Sources ELF OpenGo de Yuandong Tian et Larry Zitnick, mai 2018.

[42] Voir les explications de DeepMind dans AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, janvier 2019 et Capture the Flag: the emergence of complex cooperative agents, mai 2019. StarCraft est un jeu dynamique où l'information disponible est incomplète où il faut associer une planification à long terme et jeu temps réel, le tout avec un grand nombre d'éléments à intégrer et une combinatoire énorme associée. AlphaStar était alimenté par des parties existantes de joueurs humains, le tout étant complété par un apprentissage par renforcement. L'entraînement de l'IA a duré plusieurs semaines sur 16 TPU de Google. L'IA de DeepMind a gagné contre deux joueurs professionnels de StarCraft II. Quake est un jeu où il faut coopérer avec des joueurs humains et artificiels.

[43] Voir AlphaStar: Grandmaster level in StarCraft II using multi-agent reinforcement learning, octobre 2019.

[44] Voir OpenAI Five.

[45] Mais en exploitant un millier de PC plus des douzaines de serveurs dotés de GPGPU calculant pendant plusieurs mois et consommant 2,8 GWh d'électricité.

[46] Voir à ce sujet Reinforcement learning from scratch, par Emmanuel Ameisen, mai 2019.

[47] Voir Evolving simple programs for playing Atari games par Dennis Wilson, 2018 (8 pages). En pareille occasion, les algorithmes évolutionnaires battent les IA à base de deep learning.

En 2019, l'IA Five d'OpenAI gagnait au jeu d'esport **Dota 2** face à cinq joueurs champions du jeu[48] et une agence creative et technologique, **AKQA**, a créé une IA à base de deep learning qui a inventé un sport composite, Speedgate, s'inspirant des règles de sports existants[49].

Autre facteur de développement de l'IA, l'**Internet grand public** qui a créé de nouveaux besoins comme les moteurs de recherche et aussi permis le déploiement d'architectures massivement distribuées. L'Internet a aussi permis l'émergence de méthodes de travail collaboratives dans la recherche et les développements de logiciels, en particulier dans l'open source. Il a aussi fait émerger les fameux GAFA, ces acteurs dominants du Web grand public qui sont aussi très actifs dans l'IA. Le phénomène est en train de se démultiplier en Chine avec les fameux BATX et leur capacité à accumuler des données sans trop se soucier des règles sur la vie privée comme le RGPD européen.

Cela a généré la disponibilité de très **gros volumes de données**, via les usages de l'Internet et des mobiles, des objets connectés ou de la génomique, exploitables par différentes méthodes de machine learning et de deep learning. Le travail des chercheurs et développeurs est facilité par la publication de jeux de données ouverts (open data) pour le machine et le deep learning avec de nombreuses bases d'images et de textes disponibles pour réaliser des benchmarks. C'est le cas de la base ImageNet, de la base d'écriture manuscrite MNIST et de la base linguistique WordNet (en anglais). Ces bases sont généralement d'origine américaine et proviennent le plus souvent de grandes Universités.

Citons aussi l'**appel d'air** généré par la robotique, la conquête spatiale (Curiosity, Philae…), les véhicules à conduite assistée ou autonome, la sécurité informatique, ainsi que la lutte contre la fraude et les escroqueries en ligne. Sans compter l'objectif des japonais de s'occuper de leurs seniors avec des robots faute d'accepter politiquement une immigration de travail, ce qui a d'ailleurs changé récemment.

La diffusion de l'IA a aussi bénéficié de la culture de l'**open source** qui domine les outils de développement de solutions d'IA et les jeux de données sus-cités. Les chercheurs doivent publier des exemples de codes sources pour illustrer leurs méthodes, sur Github qui peuvent alors être reproduits et vérifiés par la communauté des chercheurs et développeurs. Ce processus permet une diffusion rapide des nouveautés algorithmiques, particulièrement autour des réseaux de neurones et du deep learning. C'est par exemple le cas des réseaux de neurones génératifs qui colorient automatiquement les photos chez Google ou ceux qui font de l'upscaling de vidéos chez Samsung.

Tout cela a abouti à la création de **nombreuses applications commerciales** de l'IA mêlant le machine learning, les objets connectés, la mobilité et le big data. Avec des attentes fortes dans le marketing, le e-commerce, la finance et le vaste secteur de la santé. Comme les usages de l'IA sont bien concrets et qu'ils touchent presque toutes les industries, on peut affirmer sans trop risquer de se tromper que la tendance est solide.

Connexionisme et symbolisme

Comme tout domaine scientifique complexe, l'IA n'a jamais été un terrain d'unanimité et cela risque de perdurer. Diverses écoles de pensée se disputent sur les approches à adopter[50].

---

[48] Voir OpenAI's Dota 2 AI steamrolls world champion e-sports team with back-to-back victories par Nick Slatt, avril 2019. Five utilise une technique d'apprentissage par renforcement dans une version virtuelle du jeu. Le jeu est plutôt complexe avec plus d'une centaine de personnages et artefacts.

[49] Voir AKQA says it used AI to invent a new sport called Speedgate par Anthony Ha, avril 2919.

[50] Voir l'excellent La revanche des neurones - L'invention des machines inductives et la controverse de l'intelligence artificielle par Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, novembre 2018 (38 pages) qui relate l'histoire de cette rivalité.

On a vu au départ s'opposer :

- Les partisans du **connexionnisme** – utilisant le principe du machine learning, du biomimétisme, des réseaux de neurones et de l'apprentissage, les réseaux de neurones étant pour l'instant surtout utilisés pour les sens artificiels (vision, parole). C'est le raisonnement inductif qui réalise des prévisions et des généralisations à partir d'observations. Ces méthodes sont le plus souvent probabilistes.
- Les partisans du **symbolisme** qui préfèrent utiliser des concepts de plus haut niveau sans chercher à les résoudre via des procédés de biomimétisme. L'IA symbolique modélise le raisonnement logique et représente les connaissances avec des objets et des symboles formels les associant entre eux (appartient à, fait partie de, est équivalent à) via d'éventuelles contraintes. C'est un raisonnement déductif qui s'appuie sur la logique reposant sur des faits et règles connus. On l'utilise par exemple pour résoudre des problèmes de mathématiques mais aussi des problèmes d'optimisation divers.

J'illustre ces deux courants avec la manière dont on apprend aux enfants en bas âge à traverser la rue. La méthode connexionniste et inductive consisterait à les laisser y aller au hasard en espérant qu'ils apprennent de leurs erreurs et ne se fassent pas écraser et que les adultes découvrent par l'observation les erreurs des enfants ou que les enfants apprennent par eux-mêmes. Ce serait du connexionnisme avec apprentissage par essais-erreurs ou par renforcement.

La méthode symbolique tire parti du savoir accumulé et des règles connues pour traverser la rue. On transmet aux enfants ces règles consistant à regarder des deux côtés de la route, à traverser sur les passages piétons, à respecter les feux et, éventuellement à traverser vite et à tenir la main de l'adulte accompagnant. Cela permet de gagner du temps et de sauver des vies !

Tout cela pour dire que si on connaît déjà les règles pour résoudre un problème ou faire une prévision, il vaut mieux les appliquer directement. On utilise la logique connexionniste lorsque l'on ne peut pas modéliser un système complexe avec des règles établies ou bien, lorsque ces règles changent très souvent et rapidement, comme dans le cas de la fraude, affectée par des fraudeurs toujours très inventifs.

L'espèce humaine passe ainsi son temps à alterner intelligence connexioniste pour comprendre le monde et le tester et intelligence symbolique pour raisonner avec les connaissances acquises.

Cette dichotomie était incarnée par la **joute intellectuelle entre "neats" et "scruffies"**, les premiers, notamment John McCarthy (Stanford), considérant que les solutions aux problèmes devraient être élégantes et carrées, et les seconds, notamment Marvin Minsky (MIT) que l'intelligence fonctionne de manière plus empirique et pas seulement par le biais de la logique[51]. Comme si il y avait un écart entre la côté Est et la côte Ouest[52] !

Le schéma *ci-dessous* cite les protagnonistes les plus connus des années 1950 à 1970. On pourrait ajouter évidemment les grands chercheurs de l'IA connexionniste comme David Rumelhart, Geoff Hinton et Yann Le Cun[53].

Ces débats ont leur équivalent dans les sciences cognitives, dans l'identification de l'inné et de l'acquis pour l'apprentissage des langues. **Burrhus Frederic Skinner** est à l'origine du comportementalisme linguistique qui décrit le conditionnement opérant dans l'apprentissage des langues.

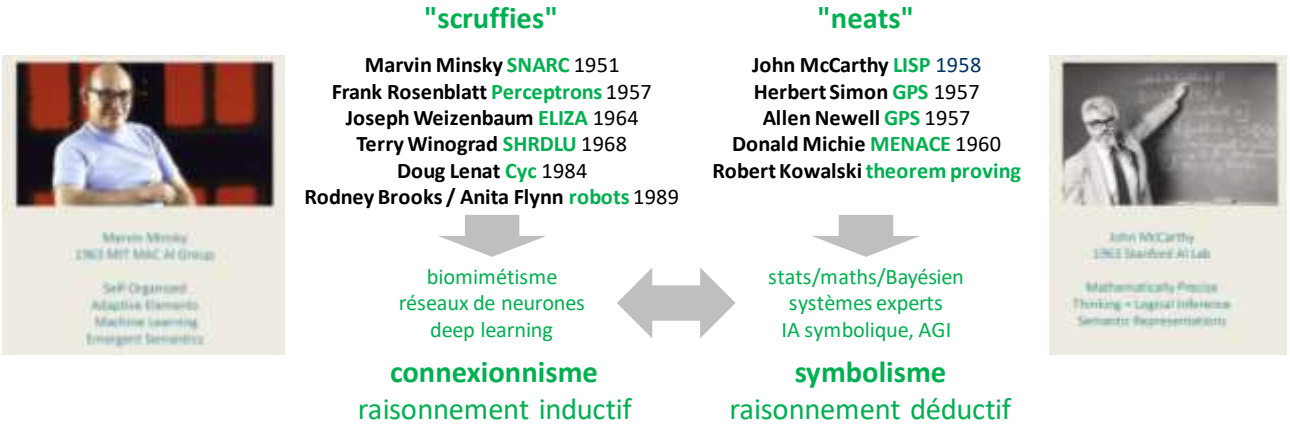
---

[51] Voir Marvin Minsky said "Probabilistic models are dead ends." What approach are researchers taking to solve the problem of Artificial General Intelligence?, 2016 et Marvin Minsky: AI "has been brain dead since the 1970s", 2003.

[52] Cette évolution entre IA symbolique et connexionniste dans le temps est analysée dans We analyzed 16,625 papers to figure out where AI is headed next par Karen Hao, janvier 2019.Voir aussi ce débat envenimé sur Twitter entre Gary Marcus et Yann LeCun en décembre 2019.

[53] J'orthographie son nom en version française en séparant le Le du Cun. A l'étranger et aux USA, Yann LeCun s'écrit d'un seul mot pour éviter que le Le soit perçu comme le middlename.

Noam Chomsky avait remis en cause cette approche en mettant en avant l'inné, une sorte de préconditionnement du cerveau des enfants avant leur naissance qui leur permet d'apprendre facilement les langues.



En gros, le fonctionnement de l'intelligence humaine est toujours l'objet de désaccords scientifiques ! On continue d'ailleurs à en découvrir toujours plus sur la neurobiologie et le fonctionnement du cerveau.

D'autres débats ont cours entre les langages de programmation déclaratifs et les moteurs d'inférences utilisant des bases de règles. Sont arrivées ensuite les méthodes statistiques s'appuyant notamment sur les réseaux bayésiens, les modèles de Markov et les techniques d'optimisation.

Après une dominance des méthodes mathématiques et procédurales, ce sont les réseaux de neurones et l'apprentissage profond les utilisant qui ont pris le dessus au début des années 1990 puis à partir de 2012, en particulier pour la vision artificielle, la reconnaissance et le traitement du langage.

La technique la plus remarquable étant celle des réseaux de neurones convolutifs, créée par le français **Yann Le Cun** et largement améliorée depuis par nombre de chercheurs, faisant elle-même suite aux TDNN (time delay neural networks) d'**Alexandre Waibel**[54]. Nous y reviendrons.

Pedro Domingos, l'auteur de « The Master algorithm », décompte en fait cinq grands courants dans l'IA en plus du symbolisme et du connexionnisme (*ci-dessous*).

Tribe	Origins	Problem	Master Algorithm
Symbolists	Logic, philosophy	Knowledge composition	Inverse deduction
Connectionists	Neuroscience	Credit assignment	Backpropagation
Evolutionaries	Evolutionary biology	Structure discovery	Genetic programming
Bayesians	Statistics	Uncertainty	Probabilistic inference
Analogizers	Psychology	Similarity	Kernel machines

Il faut ajouter celui des évolutionnistes avec les algorithmes génétiques (dont nous reparlerons), celui des bayésiens avec une vision probabiliste des choses et celui des analogistes et leurs algorithmes de clustering. Dans de nombreux cas, ces approches sont combinées pour générer des solutions optimales.

D'un point de vue historique, la vague symbolique a dominé l'IA entre ses débuts et les années 1980. Nous sommes actuellement en pleine vague connexionniste depuis l'essor fulgurant du deep learning depuis 2012.

[54] Les chercheurs français ont aussi fait avancer le domaine. Voir par exemple : Experiments with Time Delay Networks and Dynamic Time Warping for Speaker Independent Isolated Digits Recognition, de Léon Bottou, Françoise Soulié, Pascal Blanche, Jean-Sylvain Lienard, 1989 (4 pages) qui décrit une méthode de reconnaissance vocale de chiffre indépendante du locuteur.

Ce phénomène de vases communicants a eu un impact sur la capacité des IA à raisonner. On n'a pas fait de grands progrès dans ce domaine ces deux dernières décennies. En pratique, de nombreux chercheurs ambitionnent de fusionner les approches symboliques et connexionnistes pour gérer du raisonnement automatique[55]. D'autres comme DeepMind essaient de gérer toute forme de raisonnement avec des réseaux de neurones.

Définitions et segmentations

Lorsque j'ai commencé à m'intéresser sérieusement à l'IA à partir de 2015, j'ai été confronté à une difficulté : comment segmenter ce champ scientifique bien touffu ? J'ai consulté de nombreuses sources d'informations qui ne m'ont pas satisfait. Certaines comprenaient même de graves erreurs, ce dont je ne me suis rendu compte plus tard, après les avoir propagées dans mes premiers écrits sur le sujet.

L'IA est un ensemble de techniques permettant de résoudre des problèmes complexes en s'inspirant de mécanismes cognitifs humains, agissant de manière rationnelle en fonction de faits, données et expériences, et capables d'atteindre de manière optimale un ou plusieurs objectifs donnés.

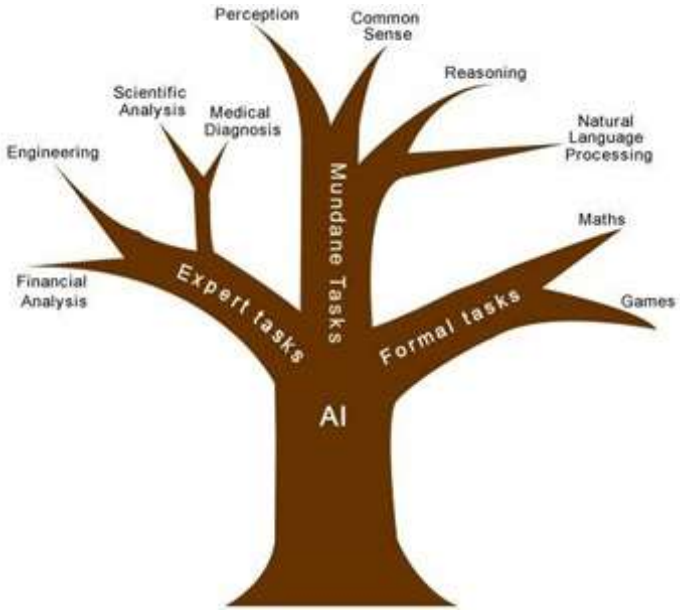
La rationalité n'est pas l'omniscience mais la capacité à agir en fonction des informations disponibles, y compris celles qui sont ambigües. Cette rationalité est habituellement limitée par notre volonté, le poids émotionnel de notre cerveau limbique et notre capacité d'optimisation.

A haut niveau, on peut découper l'IA en trois grands domaines, dont deux que nous avons déjà décrits précédemment :

- Le **symbolisme** qui se focalise sur la pensée abstraite et la gestion des symboles, l'algorithmique et la logique. Le symbolisme modélise notamment les concepts sous la forme d'objets reliés entre eux par des prédicats logiques (appartient à, etc). C'est une approche « macro » de résolution de problèmes. C'est dans cette catégorie que l'on peut ranger les systèmes experts et moteurs de règles qui les font fonctionner, et dans une certaine mesure, le web sémantique. Intuitivement, on peut considérer que la connaissance humaine est le résultat de l'accumulation d'observations obtenues par des méthodes connexionnistes. Il peut en être autant dans le couplage entre l'IA symbolique et l'IA connexionniste.
- Le **connexionnisme** qui se focalise sur la perception, dont la vision, la reconnaissance des formes et s'appuie notamment sur les réseaux neuronaux artificiels qui reproduisent à petite échelle et de manière approximative le fonctionnement générique du cerveau. C'est une vision « micro » et probabiliste de la résolution des problèmes. C'est ici que l'on peut ranger le deep learning utilisé dans la vision artificielle ou le traitement de la parole. Cette IA est aussi associable aux méthodes stochastiques et heuristiques du machine learning. L'IA connexionniste s'entraîne avec des données et/ou en interagissant avec son environnement via le principe de l'entraînement par renforcement.
- Le **comportementalisme** qui s'intéresse aux pensées subjectives de la perception. C'est dans ce dernier domaine que l'on peut intégrer l'informatique affective (ou affective computing) qui étudie les moyens de reconnaître, exprimer, synthétiser et modéliser les émotions humaines. C'est une capacité qu'IBM Watson était censé apporter au robot Pepper de Softbank Robotics (ex Aldebaran). Dans la pratique, cette troisième catégorie est un peu hors sol et peut être mise en œuvre avec un mix d'IA symbolique et connexionniste.

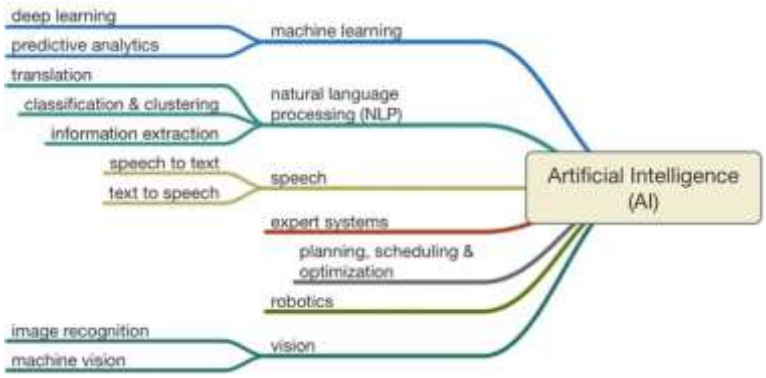
[55] Voir Que devient l'IA symbolique, Olivier Ezratty, en avril 2018.

Reprenant plus ou moins ce découpage, cette segmentation sous forme d'arbre comprend trois grandes branches : l'une pour les **tâches d'expertise**, la seconde pour les **tâches courantes** (perception, sens commun, raisonnement, langage) et la troisième pour les **tâches formelles** (jeux, mathématiques).

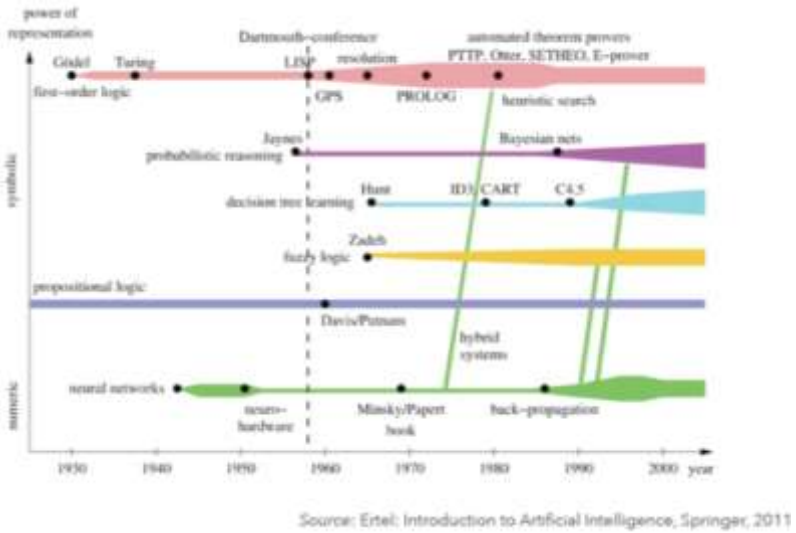


Cette autre segmentation très utilisée aux alentours de 2015 (*schéma suivant*) comprend le machine learning, le deep learning, le traitement du langage, les systèmes experts, la robotique et la vision.

Elle place curieusement au même niveau des outils génériques comme le machine learning et le deep learning et ses propres applications comme la vision artificielle ou le traitement du langage. La robotique intègre de son côté tous les autres champs du schéma plus quelques autres qui lui sont spécifiques comme les capteurs, les matériaux, la mécanique, les moteurs électriques et autres batteries.



Cette autre segmentation datant de 2011, donc d'avant l'explosion cambrienne du deep learning, illustre l'évolution des grandes techniques de l'IA symbolique (en haut) et connexionniste (en bas) lors des 50 premières années de son Histoire[56].

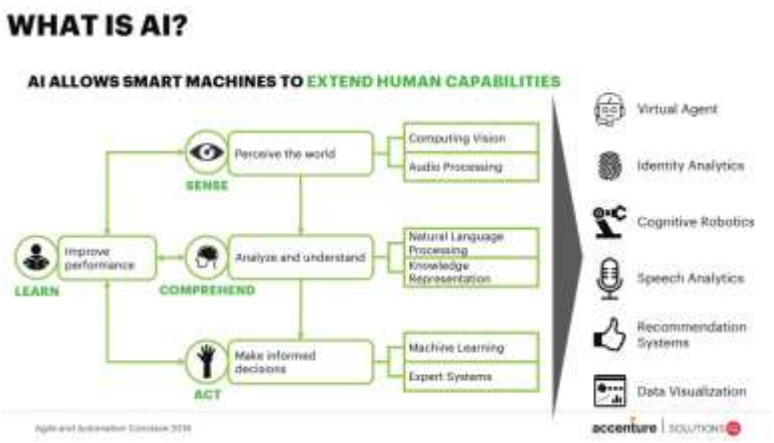


Source: Ertel: Introduction to Artificial Intelligence, Springer, 2011

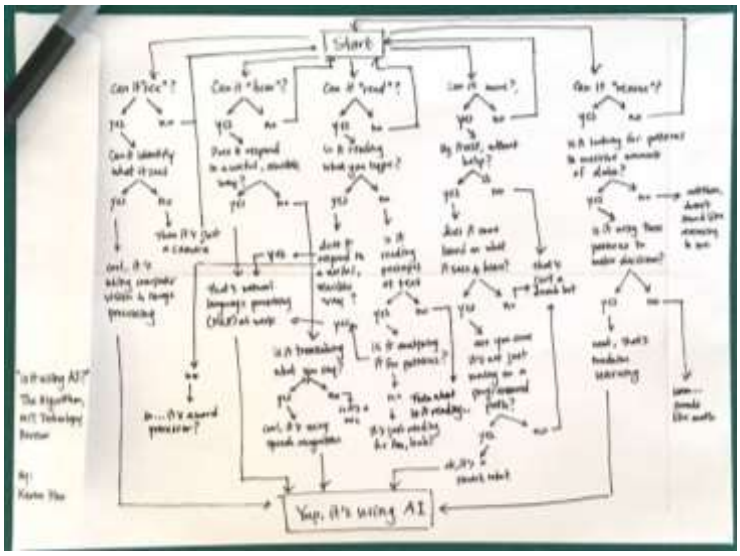
Elle rappelle aussi que ces différents courants de recherche se nourrissent les uns les autres comme dans n'importe quelle discipline scientifique. Il faudrait que cela puisse continuer !

[56] Voir aussi cet article intéressant : An intuitive, high-level framework to understand the technical trends in Artificial Intelligence par Charles Vr, septembre 2017. Il décrit bien les différences d'exigences entre les tâches de perception et celles de décision.

Celle-ci qui provient d'une présentation d'Accenture de 2018[57] fait un tri assez élégant entre techniques liées à la perception, à l'analyse et à la prise de décision. Elle associe dans les deux dernières catégories les briques de l'IA connexionniste (NLP, machine learning) et de l'IA symbolique (représentation des connaissances, systèmes experts). Il manque les réseaux multi-agents qui jouent un rôle important dans l'intégration des briques de l'IA.



Voici une autre forme de segmentation, assez réductrice, qui reprend en quelque sorte la précédente, et passe par un arbre de décision pour déterminer si une solution relève de l'IA[58]. Il faut en gros qu'elle voie, lise, entende, se déplace (robot) ou raisonne. Cela exclut cependant de nombreuses applications de l'IA, au niveau des analyses prédictives, de segmentation automatique, de réseaux multi-agents. Cette arborescence est surtout adaptée aux fonctions de perception : vision et traitement du langage.



Ici [59], les approches symboliques sont intégrées dans l'apprentissage automatique, ce qui n'est pas forcément exact. Les règles qui alimentent les moteurs de règles sont souvent captées manuellement. IBM Watson est en effet parfois utilisé pour extraire des règles par traitement du langage qui alimentent ensuite des moteurs de règles. Un filtre humain est nécessaire ce qui enlève la notion d'automatique à l'apprentissage. Et qu'est-ce qui est de l'IA et ne relève pas de l'apprentissage automatique ? Les réseaux multiagents ou la programmation par contraintes ? On n'en saura rien selon ce schéma !

[57] Voir AI in Agile, AI and automation conclave, Anubhav Gupta, 2018, Accenture (34 slides).

[58] Voir Is this AI, we drew a flowchart to work it out par Karen Hao dans MIT Technology Review, novembre 2018.

[59] Voir Regards croisés sur l'IA, Grand Lyon 2019 (36 pages).

Le rapport **France IA** publié en mars 2017 par le gouvernement français proposait pour sa part une segmentation plus fouillée, compilant les principaux travaux de recherche du domaine en France. Mais cela reste encore touffu[60].

Bref, il faudrait inventer une IA qui segmente convenablement le champ de l'IA ! J'en propose une page 33 qui comprend le raisonnement automatique, le machine learning et les réseaux multi-agents.

A un niveau conceptuel plus elevé, voici une segmentation qui relie entre eux quatre domaines de manière plus hiérarchique (cf schéma *ci-dessous*) :

- Les **solutions** : que l'on va directement utiliser dans les entreprises ou chez les particuliers avec les chatbots, les véhicules autonomes, les robots, les systèmes de recommandation, les outils de segmentation client, le marketing prédictif ou les solutions de cybersécurité.
- Les **outils** : qui aident à créer ces solutions, comme la vision artificielle, la reconnaissance de la parole, la traduction automatique, les systèmes experts, les outils de prévision ou de segmentation automatiques. J'y ai ajouté les réseaux multi-agents qui coordonnent l'action des différents outils d'une solution d'IA.
- Les **techniques** : sur lesquelles sont construits ces outils, avec les méthodes de machine learning, les réseaux de neurones, les nombreuses méthodes de deep learning et les moteurs de règles.
- Les **données** : les sources de données correspondantes et les **capteurs** associés qui jouent un rôle indispensable, notamment pour les approches connexionnistes, le machine learning et le deep learning.

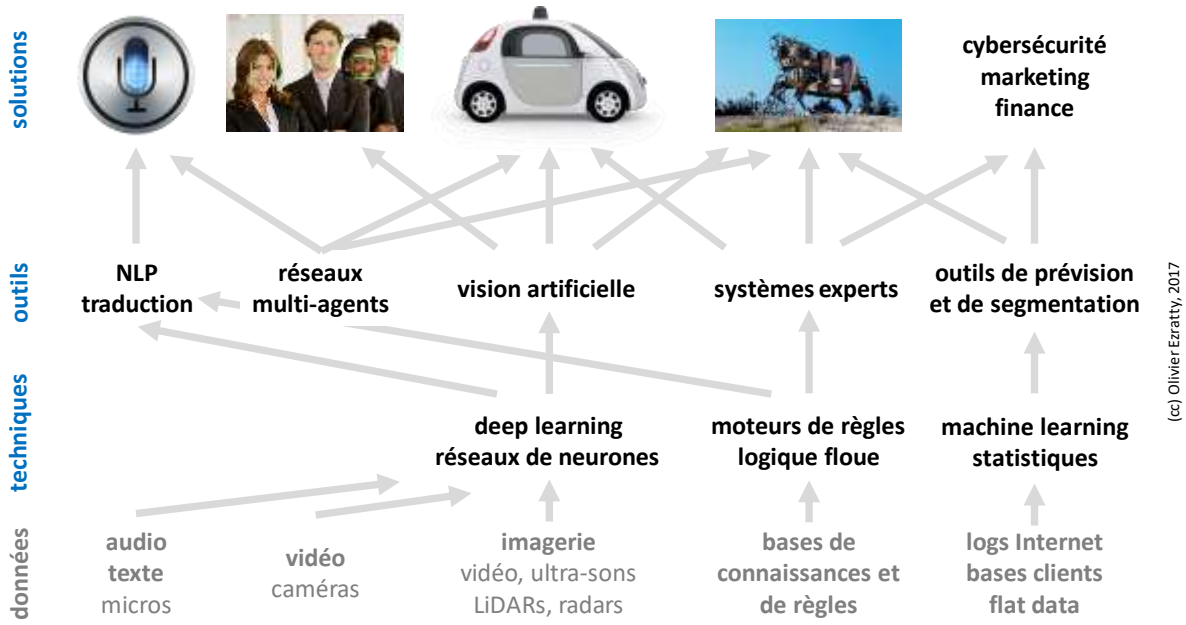
Cela rappelle que les solutions à base d'IA sont des assemblages de diverses briques logicielles et matérielles selon les besoins. Ces briques sont des plus nombreuses. A tel point que leur intégration est un enjeu technique et métier de taille, peut-être le plus complexe à relever[61].

Lorsqu'une startup indique qu'elle a créé « une IA » pour faire ceci ou cela, cela signifie qu'elle a assemblé des techniques, paramétré des outils, en général assez standards, pour exploiter des données, et les a appliqués pour créer une solution. L'originalité est rarement dans la création des briques mais plutôt dans leur sélection, leur combinaison, leur assemblage et le problème métier traité.

---

[60] Voir Stratégie France I.A. : Pour le développement des technologies d'intelligence artificielle, mars 2017.

[61] Aymeric Poulain Maybant évoque sa thèse de doctorat dans L'Intelligence Artificielle en questions - Une spécialité qui se cherche en 2014. Il évoque l'hybridation en sciences cognitives qui date de 2005 et décrit très bien cet enjeu. L'IA intégrative est un des principaux facteurs de développement du secteur. On le retrouve dans l'association de nombreuses techniques dans les solutions d'IA comme le couplage de réseaux neuronaux et d'approches statistiques plus simples, notamment dans la reconnaissance de la parole.



Limites de l'IA

Je vous livre ici le fruit d'une réflexion permettant de faire le tri de différentes méthodes qui relèvent de l'IA et de celles qui n'en sont pas. Le schéma *ci-dessous* illustre cela.

Les deux premières colonnes positionnent l'IA connexionniste et l'IA symbolique. La première extrait des règles de données observées et généralement taggées par l'Homme. Elle permet d'identifier de manière empirique des règles, par exemple sur le comportement de clients. Ces règles peuvent à leur tour alimenter une IA symbolique qui exploite des faits et règles connus et formalisés pour résoudre des problèmes de logique. En quelque sorte, le machine learning est une brique d'alimentation du raisonnement automatique[62].

La troisième colonne positionne l'interaction entre l'IA et le monde réel. C'est ce qui lui permet de faire de l'apprentissage par renforcement. Elle vise à évaluer la réaction du monde réel à des actions externes. C'est le cas d'un chatbot qui évalue la qualité de son dialogue et de ses réponses. C'est celui du robot qui apprend par tâtonnement à maîtriser ses gestes, comme un enfant en bas âge qui apprend à marcher. Hors IA, c'est aussi le champ de l'expérimentation biologique en boîtes de Petri. Ces interactions s'appuient sur des mécaniques tellement complexes qu'elles ne sont pas simulables « in silico ». L'expérimentation avec le monde réel permet d'en extraire des règles qui peuvent à leur tour également alimenter de l'IA symbolique. Par exemple : telle protéine agit de telle manière sur tel type de cellule à telle température.

La dernière colonne positionne les situations et problèmes qui sont simulables ou partiellement simulables « in silico », dans des ordinateurs, sans passer par le monde réel et l'expérimentation de la colonne précédente.

On y trouve deux catégories de problèmes : ceux qui s'expriment avec des règles simples et discrètes et dont la simulation relève souvent de l'élagage d'un arbre de décision. C'est le cas des jeux de société comme le jeu d'échecs et celui de Go. Et puis celui de la physique et du monde biologique qui s'appuie sur des règles complexes, des équations différentielles ou linéaires, continues, et dont la simulation commence à être possible mais est difficile. Cela concerne par exemple la simulation du repliement des protéines dans les cellules, un des problèmes les plus complexes à résoudre qui soit pour de grandes protéines[63].

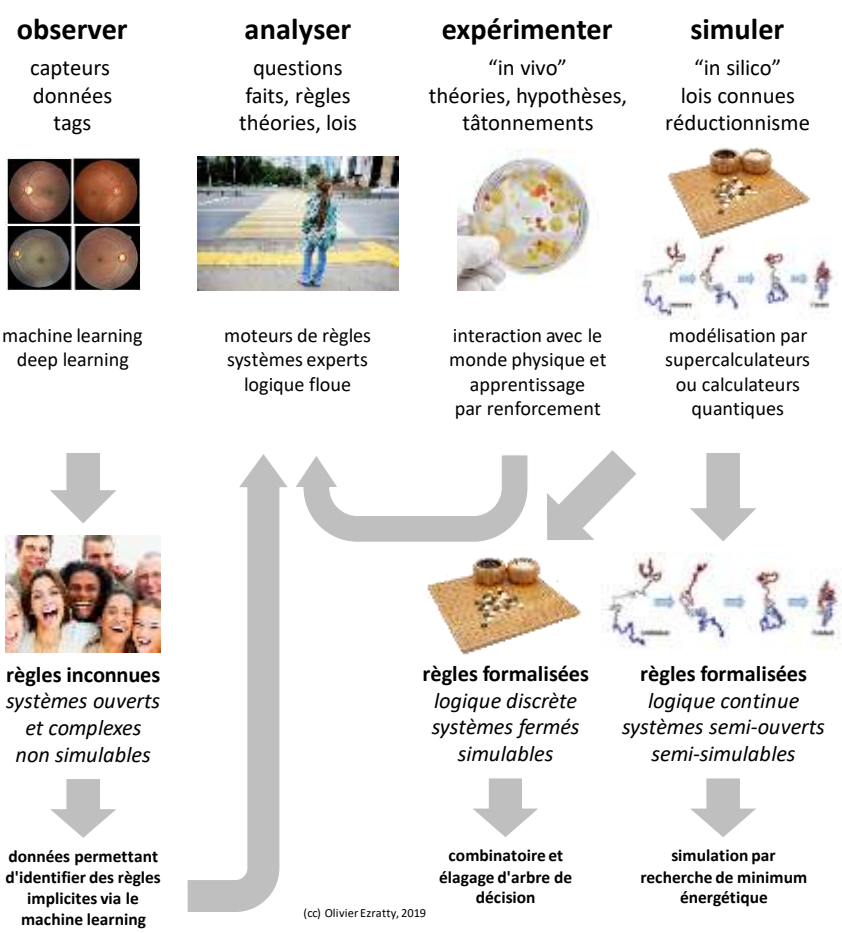
---

[62] Voir AI is about machine reasoning – or why machine learning is just an fancy plugin, de la startup allemande Arago, 2017.

[63] Ce problème est actuellement résolu de manière approximative avec du deep learning (apprentissage sur l'existant). D'ici quelques décennies, lorsque les ordinateurs quantiques seront assez puissants, on passera à la simulation qui sera plus précise et plus puissante.

C'est l'un des champs d'application du calcul quantique. Lorsque les règles sont discrètes comme avec le jeu de Go, la simulation s'effectue avec des réseaux de neurones et fait donc partie du champ de l'IA. Lorsque la simulation relève de règles continues et mathématiques, et qu'on les modélise avec du calcul scientifique ou du calcul quantique, on sort du champ de l'IA.

Il en va de même lorsque l'on réalise de la simulation de systèmes complexes par approximation, comme avec l'usage de la méthode des éléments finis pour des simulations d'écoulement d'air sur les ailes d'avions ou pour des simulations météo. On alimente les simulations avec des états réels mesurés, comme les images satellites et données de terrain pour la météo.



On commence à parler de jumeau numérique pour représenter un système et effectuer des simulations dessus au lieu de faire les expérimentations dans le monde réel, comme pour les crash-tests de véhicules ou la simulation de vol d'un aéronef. Cette dernière colonne correspond aux limites de l'IA et des techniques d'origine humaine.

Elles sont liées aux théories de la complexité qui régissent la nature des problèmes que l'on peut résoudre ou pas avec des calculs[64].

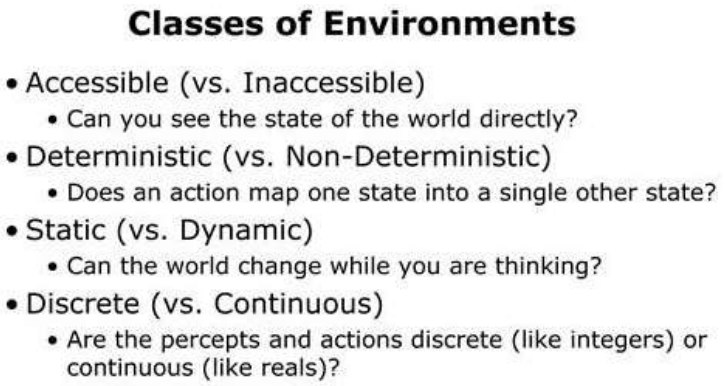
Dans la seconde et la quatrième colonne, des problèmes de logique ou de simulation sont tellement complexes qu'ils sont indécidables. Ils ne peuvent pas avoir de solution. C'est pour cela que l'IA ne prévoit généralement pas les « cygnes noirs » et que la société ne peut pas se mettre facilement en équation[65]. On en est alors réduit à faire du réductionnisme, en réduisant par grandes approximations des problèmes complexes à un ensemble de problèmes plus simples. Et on est obligé de faire des expériences.

Le scénario qui le reflète le mieux est celui de la détermination d'une politique économique d'un gouvernement. Idéalement, il faudrait mettre l'économie d'un pays et du monde en équations et simuler la modification de paramètres clés pour évaluer différentes politiques. D'un point de vue opérationnel, le nombre de paramètres est tellement important et reposant sur des dimensions sociales et psychologiques complexes, il est impossible de simuler l'ensemble. On en est réduit à opérer par tâtonnement, ou en s'accrochant aux branches d'idéologies économiques établies, et peut-être aussi à accepter que le monde n'est pas déterministe et régit par des règles immuables !

[64] Voir la partie dédiée aux théories de la complexité de ma série d'articles sur l'informatique quantique, juillet 2018.

[65] Voir à ce sujet l'intéressant Pourquoi la société ne se laisse pas mettre en équations ? de Hubert Guillaud, avril 2018.

Comme le décrit la liste *ci-contre*, les paramètres d'un système à simuler ou à expérimenter sont nombreux. Est-ce que le système est accessible ou pas pour être mesurable ? Est-ce qu'il est déterministe, à savoir que des actions génèrent les mêmes conséquences à chaque itération ? Est-ce que le système est dynamique ou statique ? Fonctionne-t-il en vase clos (comme un jeu de Go) ou dans un monde ouvert (comme pour une politique économique) ?



Est-ce que les règles de fonctionnement sont discrètes (oui/non) ou continues (régies par des nombres réels) ? Tous ces paramètres vont conditionner la faisabilité et le réalisme des expérimentations et des simulations pour résoudre un problème ou prédire un état futur d'un système complexe.

Pour reboucler sur le schéma, la première colonne qui relève du machine learning illustre le fait que l'IA connexionniste est une manière de contourner l'impossibilité de simuler le monde physique. A la place, on l'observe et on utilise des méthodes probabilistes pour en déduire des règles empiriques et faire des prévisions approximatives. Mais la théorie du chaos rend le monde en grande partie imprédictible ! Soit-dit en passant, heureusement.

Etat des lieux

Faisons ici un point rapide sur l'état des lieux de l'IA en ce début 2021. Le schéma qui suit positionne de manière rudimentaire la comparaison des capacités des machines face à celles de l'Homme. Cette approche anthropomorphique de l'IA n'est pas la seule qui soit pertinente mais elle permet de bien remettre les choses en place.

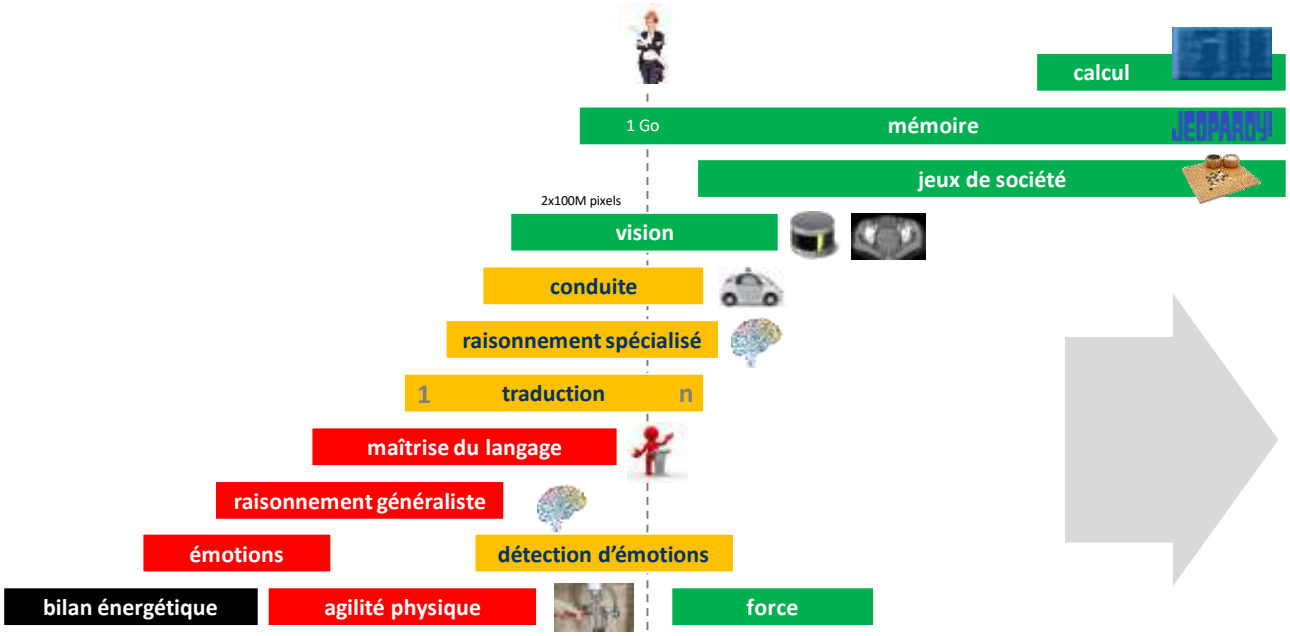
Les domaines où l'IA et les outils numériques en général dépassent déjà l'Homme sont ceux qui relèvent du calcul, comprenant les applications « data » du machine learning pour faire des prévisions ou de la segmentation, tout ce qui relève de la mémoire, notamment dans l'accès à de gros volumes d'information, les jeux de société qui sont maintenant presque tous gagnés par l'IA face à l'Homme, les derniers en date étant le Go et le Poker, et enfin, la vision artificielle, notamment celle qui est appliquée à des fonctions spécialisées comme dans le diagnostic médical.

Dans les domaines où nous sommes presque ex-aequo, on peut citer la conduite autonome dans certaines conditions et le raisonnement spécialisé.

Pour le reste, même si les progrès sont constants, les IA ont encore du chemin à faire pour atteindre les capacités humaines. Tout ce qui relève de la maîtrise du langage est encore en deça des capacités humaines. La traduction automatique dépasse les capacités générales de l'Homme mais pas celle des spécialistes. Un véritable bilingue fera toujours mieux qu'un système automatique. Quant aux agents vocaux, ils rendent des services mais sont très loin de passer le test de Turing pour égaler un spécialiste humain, même avec Google Duplex qui prend rendez-vous avec votre docteur à votre place.

Le raisonnement généraliste est encore hors de portée avec l'IA actuelle[66]. Ceux qui cherchent à le reconstituer sont obligés de le décomposer en tâches et sous-tâches et de les mettre en œuvre laborieusement.

[66] Voir à ce sujet Don't believe the hype: separating AI fact from fiction de Larry Lefkowitz, juillet 2018 qui évoque la dimension mécanique de l'IA actuelle et sa difficulté à gérer du raisonnement et de la compréhension du sens.



L'agilité générale des robots reste enfin très limitée par rapport à celle de l'Homme même si l'on peut admirer les prouesses des robots de **Boston Dynamics**, qui sont d'ailleurs en partie « fake » car ces robots sont en fait téléguidés par des opérateurs humains. C'est pour cela que l'on n'est pas près de voir des robots s'occuper des patients dans les hôpitaux ou les maisons de retraite ou que le robot de **Samsung** présenté au CES 2021 qui remplit le lave-vaisselle n'est pas encore sur le point d'être commercialisé[67].

Deux derniers points sont encore en devenir : la gestion des émotions et le bilan énergétique. Les IA n'ont pas d'émotions et d'envies, en tout cas autonomes. Elles peuvent détecter des signes extérieurs d'émotions humaines et réagir de manière mécanique à ces émotions. Pour ce qui est de l'énergie, les IA sont en général plus voraces que l'Homme avec son cerveau qui ne consomme que 20W.

En pratique, les solutions exploitant de l'IA rendent de très nombreux services au quotidien au grand public et aux entreprises, mais sans forcément relever de l'imitation de l'intelligence humaine, notamment sa capacité de raisonnement. Il en va ainsi des systèmes à base de machine learning qui essayent laborieusement de vous proposer le bon film sur Netflix ou le bon produit sur Amazon en exploitant votre historique d'usage et celui des autres Internautes.

Ces techniques de machine learning sont basées sur des méthodes et outils probabilistes qui ne correspondent pas aux sens humains. Cela permet de modérer à court terme ses frayeurs sur l'IA. Autre limitation, ils n'ont que trop rarement le moyen d'intégrer un effet de surprise dans leurs recommandations, car ils ne sont guidés que par votre consommation passée, pas par d'autres sources d'inspiration qui n'ont jamais été capturées en ligne.

Il existe d'autres segmentations de l'IA par rapport aux capacités humaines comme ces quatre catégories de systèmes : qui pensent comme des humains, qui agissent comme des humains, qui pensent rationnellement (et donc, pas forcément comme des humains) et qui agissent rationnellement (itou). On peut aussi plus simplement utiliser trois catégories : les sens, le raisonnement et l'action. C'est la combinaison des trois qui fait encore de nous les Humains des êtres différents des machines. Pourvu que cela dure !

---

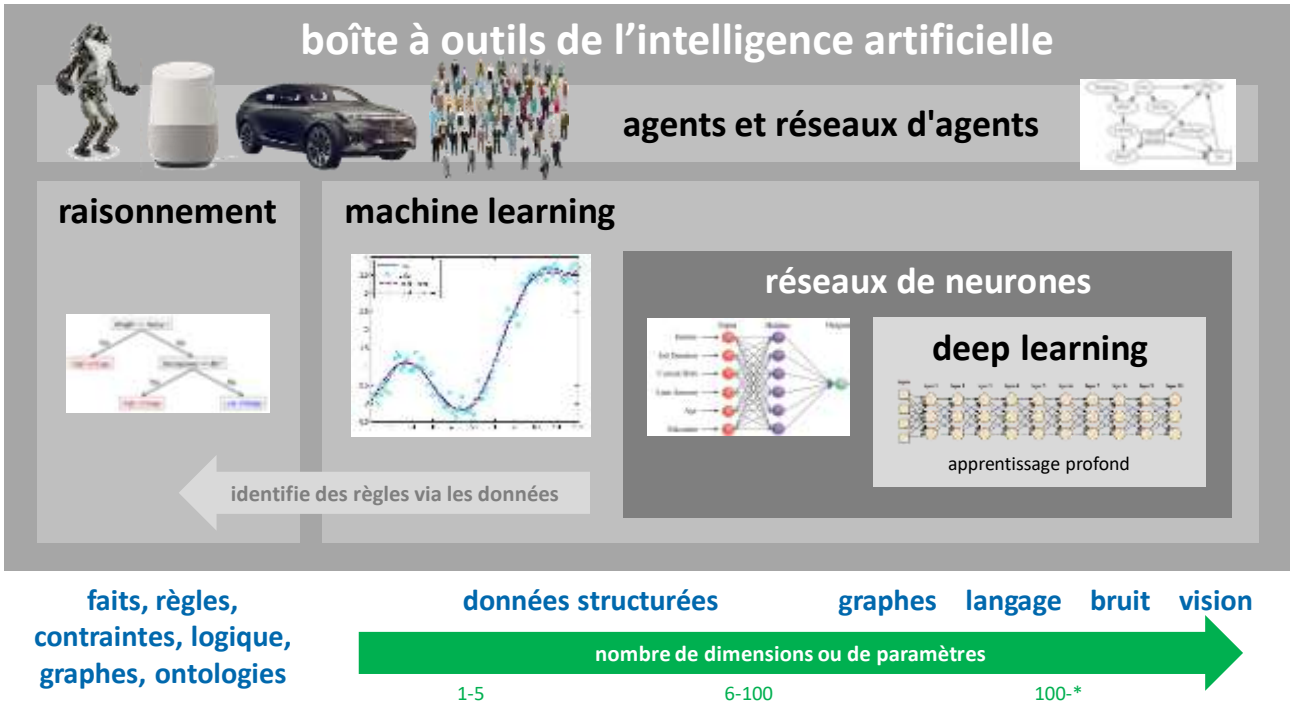
[67] Voir Progress in AI isn't as Impressive as You Might Think de Will Knight, novembre 2017 qui cite le rapport du MIT From Artificial Index Report 2017 (101 pages) piloté par Erik Brynjolfsson, le co-auteur de The Second Machine Age en 2014 et qui considère que l'on est dans une bulle de l'IA. Pour le robot de Samsung, voir Samsung's CES 2021 robots will clean your house and pour you a glass of wine par Shara Tibken dans C-NET, janvier 2021.

Algorithmes et logiciels de l'IA

Pour décrire l'univers des algorithmes, logiciels et outils de développement de l'IA, j'utilise et améliore régulièrement une segmentation simplifiée du domaine apparue dans l'édition 2017 de cet ebook avec les outils du raisonnement automatique, ceux du machine learning qui intègrent les réseaux de neurones et du deep learning et enfin, la couche d'intégration des briques de l'IA que constituent les notions d'agents et de réseaux multi-agents.

Pour chacune de ces briques fondamentales, je vais évoquer si besoin est leur ancienneté, les progrès les plus récents, les applications phares ainsi que quelques acteurs des marchés correspondants, notamment au niveau des outils de développement.

Cette segmentation couvre les principaux usages actuels de l'IA puisque l'essentiel des solutions de traitement de l'image, du langage naturel et de la perception relèvent du deep learning, celles qui gèrent des données structurées et notamment de la prévision relèvent du machine learning de base et celles qui traitent du raisonnement et de la planification relèvent de différentes variations de moteurs de règles, solveurs et outils associés.



De nombreux pans de l'IA qui ne figurent pas dans ce schéma exploitent ses différentes briques :

- Les **algorithmes évolutionnaires** ou génétiques qui peuvent s'appuyer sur du deep learning et qui testent plusieurs versions de solutions pour ne conserver que les meilleures.
- La **représentation des connaissances** qui les extrait de données textuelles non structurées, par exemple, via du deep learning, et les exploite ensuite le plus souvent dans du raisonnement automatique avec les outils de l'IA symbolique. C'est le sens de la flèche latérale grise dans le schéma qui décrit l'alimentation d'approches symboliques par des règles extraites de données via le deep learning.
- L'**IA affective** qui exploite une panoplie large d'outils du machine learning et du deep learning voire du raisonnement automatique pour capter et classifier des éléments extérieurs des émotions humaines et agir en conséquence.

- Le **transfer learning**, ou apprentissage par transfert, est une variante du deep learning qui permet d'entraîner un réseau de neurones à partir d'un réseau de neurones déjà entraîné pour le compléter, le mettre à jour ou l'utiliser dans un domaine voisin du domaine initial. Cette technique permet de fortement réduire les apprentissages doublonnés de réseaux neuronaux pour des usages identiques. La mutualisation de réseaux neuronaux et leur accès en « open data » diminuerait aussi l'impact environnemental de l'apprentissage de l'IA.

Reprenons ces grandes briques une par une :

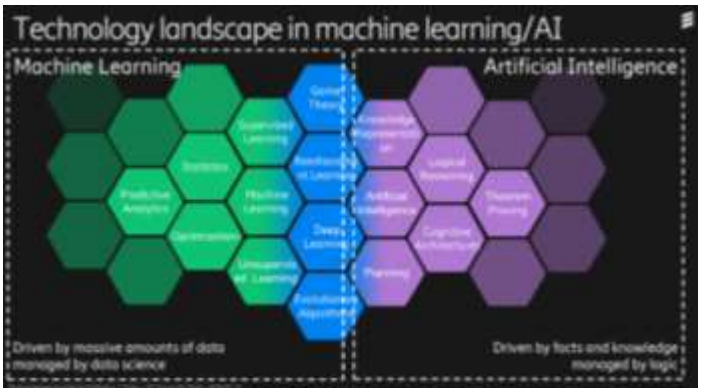
- Les **briques du raisonnement et de la planification** qui permettent de résoudre des problèmes de logique où l'on dispose d'éléments d'informations factuels sous forme de règles, faits et contraintes, dont il faut disposer au préalable pour les exploiter. Ces problèmes peuvent relever de la maintenance, de la manière de gérer la conduite d'un véhicule autonome, du fonctionnement d'un robot ou d'une machine à commande numérique dans une usine ou de l'optimisation de l'exploitation de ressources multiples pour accomplir des tâches. Les outils les plus connus dans ce domaine sont les systèmes experts bâtis avec des solveurs qui exploitent des bases de règles formelles et de faits. Les moteurs de règles s'appellent maintenant les BRMS pour Business Rules Management Systems et sont souvent intégrés dans des DMS, pour Decision Management Systems. Ces systèmes peuvent intégrer de la logique floue pour gérer des faits et règles imprécis.
- Le **machine learning** ou apprentissage automatique sert à faire des prévisions, de la classification et de la segmentation automatiques en exploitant des données en général multidimensionnelles, comme une base de données clients ou un log de serveur Internet. Le machine learning relève d'une approche probabiliste. Les outils du machine learning servent à exploiter les gros volumes de données des entreprises, autrement dit le « big data ». Le machine learning peut s'appuyer sur des réseaux de neurones simples pour les tâches complexes portant sur des données multidimensionnelles.
- Les **réseaux de neurones** constituent un sous-domaine du machine learning pour réaliser des tâches identiques, mais lorsque l'espace probabiliste géré est plus complexe. Ce biomimétisme élémentaire est exploité lorsque la dimension du problème à gérer est raisonnable. Sinon, on passe rapidement au deep learning, notamment pour le traitement de l'image et du langage.
- Le **deep learning** ou apprentissage profond, permet d'aller plus loin que le machine learning pour reconnaître des objets complexes comme les images, l'écriture manuscrite, la parole et le langage. Le deep learning exploite des réseaux de neurones multicouches, sachant qu'il en existe de très nombreuses variantes. Ce n'est cependant pas la solution à tous les problèmes que l'IA cherche à traiter[68]. Le deep learning permet aussi de générer des contenus ou d'améliorer des contenus existants, comme pour colorier automatiquement des images en noir et blanc. Deep veut dire profond. Mais le deep learning ne réfléchit pas. Il est profond parce qu'il exploite des réseaux de neurones avec de nombreuses couches de représentation intermédiaire des informations. Le deep learning n'est pas dédié exclusivement au traitement de l'image et du langage et peut aussi servir dans d'autres environnements complexes comme dans le traitement de graphes, comme en génomique ou en chimie. On l'utilise aussi dans des approches dites multimodales qui intègrent différents sens comme la vision et le langage. Enfin, le deep learning est aussi testé dans des approches symboliques, avec les Symbolic Neural Networks.
- Les **réseaux d'agents ou systèmes multi-agents** sont un domaine méconnu qui couvre la science de l'orchestration d'agents logiciels autonomes interagissant avec leur environnement. Les réseaux multiagents peuvent servir à faire de la simulation, comme celle du comportement des foules. Les agents peuvent être homogènes ou hétérogènes. L'assemblage d'agents dans des réseaux multi-agents hétérogènes est une version « macro » de la création de solutions d'IA.

[68] Voir Deep learning is not AI future, de Fabio Ciucci publié en août 2017.

Un chatbot comme un robot est toujours un assemblage hétéroclite de briques avec des moteurs de règles, du machine learning et plusieurs techniques de deep learning. Les réseaux d'agents sont donc à la fois des objets conceptuels et des outils d'assemblage de briques logicielles de l'IA.

Cette classification n'est pas la seule du marché. Il y en a quasiment autant que de spécialistes et non spécialistes du domaine de l'IA.

Celle-ci est issue d'une présentation d'**Ericsson**. Elle sépare le machine learning de l'intelligence artificielle qui est là, attachée au raisonnement, correspondant à la première case de mon schéma[69]. C'est une vision historique, pré-2012 de l'IA, réduite à l'IA symbolique, et réductrice puisque le deep learning, qui est un sous-ensemble du machine learning, réalise des tâches qui imitent des composantes de l'intelligence humaine comme la vision et la compréhension du langage.



Force brute et arbres de recherche

La force brute est l'inverse métaphorique de l'intelligence. C'est un moyen courant de simuler l'intelligence humaine ou de la dépasser. Pour un jeu comme les échecs, elle vise à tester toutes les possibilités de jeu et à identifier les chemins les plus optimums parmi des zillions de combinaisons possibles.

La force brute n'est opérationnelle que si la combinatoire à tester reste dans l'enveloppe de puissance de l'ordinateur utilisé. Si elle est trop élevée, des méthodes de simplification des problèmes et de réduction de la combinatoire doivent être utilisées. On utilise alors des algorithmes d'élagage qui évacuent les "branches mortes" de la combinatoire ne pouvant aboutir à aucune solution. C'est là qu'intervient une vague forme d'intelligence, mais qui repose sur une véritable force brute tout de même. Elle est systématique et pas intuitive. Mais la frontière entre les deux risque de s'atténuer.

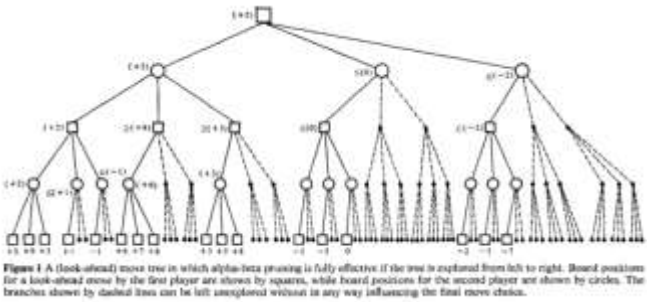
C'est d'ailleurs plus facile à réaliser aux échecs qu'au jeu de Go. La combinatoire du premier est plus faible que celle du second, notamment du fait de la taille de la grille de jeu qui est respectivement de 8x8 et 19x19 cases (*ci-contre*) !

Cela explique pour l'IA a gagné au Go après l'avoir fait aux échecs !

La force brute a été notamment utilisée par l'ordinateur **Deep Blue** d'IBM pour gagner aux échecs en 1996 puis en 1997 contre Gary Kasparov, en évaluant 200 millions de positions par seconde grâce à 510 processeurs travaillant simultanément.

[69] Voir Ericsson Press Seminar @MWC2018, février 2018 (12 slides).

IBM Deep Blue (1996)
200 millions de positions testées par secondes
510 processeurs

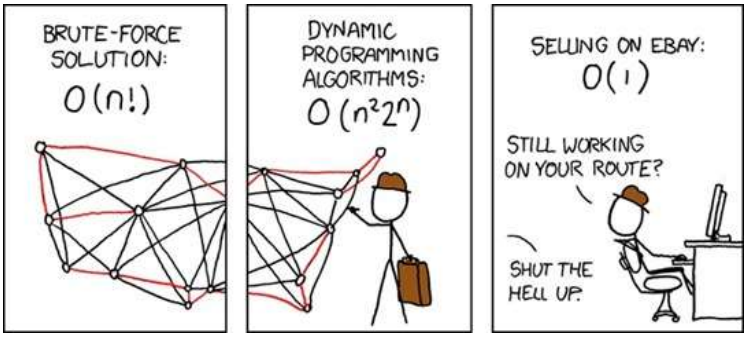


En 2016, des réseaux de neurones étaient exploités pour gagner au jeu de Go avec AlphaGo de **DeepMind**, la filiale d'IA de Google. AlphaGo exploite ainsi un mélange de force brute et de deep learning permettant de faire des économies de combinatoires à tester pour identifier les meilleurs coups.

La combinatoire du jeu de Go est en effet de plusieurs ordres de grandeur supérieure à celle des échecs. La première version d'AlphaGo bénéficiait aussi d'un apprentissage supervisé par l'exploitation de parties de Go existantes, et d'un apprentissage par renforcement, le système apprenant en jouant contre lui-même. Dans sa dernière version en octobre 2017, AlphaGo Zero élague son arbre de recherche de positions avec une architecture plus simple et unifiée à base de réseaux de neurones récursifs[70].

La force brute est utilisée dans de nombreux domaines comme dans les moteurs de recherche ou la découverte de mots de passe simples. On peut considérer que de nombreux pans de l'IA l'utilisent, même lorsqu'ils s'appuient sur des techniques modernes de deep learning ou de machine learning que nous traiterons plus loin.

La force brute s'est aussi généralisée parce que la puissance des ordinateurs le permet : ils tournent plus vite, sont distribuables, le stockage coûte de moins en moins cher, les télécommunications sont abordables et les capteurs de plus en plus nombreux, des appareils photo/vidéo des smartphones aux capteurs d'objets connectés divers[71].



Méthodes statistiques

Les méthodes statistiques et notamment bayésiennes permettent de prévoir la probabilité d'événements en fonction de l'analyse d'événements passés.

Les réseaux bayésiens utilisent des modèles à base de graphes pour décrire des relations d'interdépendances statistiques et de causalité entre facteurs (exemple *ci-dessous*).

Les applications sont nombreuses comme la détection de potentiel de fraudes dans les transactions de cartes bancaires ou l'analyse de risques d'incidents pour des assurés. Elles sont aussi très utilisées dans les moteurs de recherche au détriment de méthodes plus formelles[72].

---

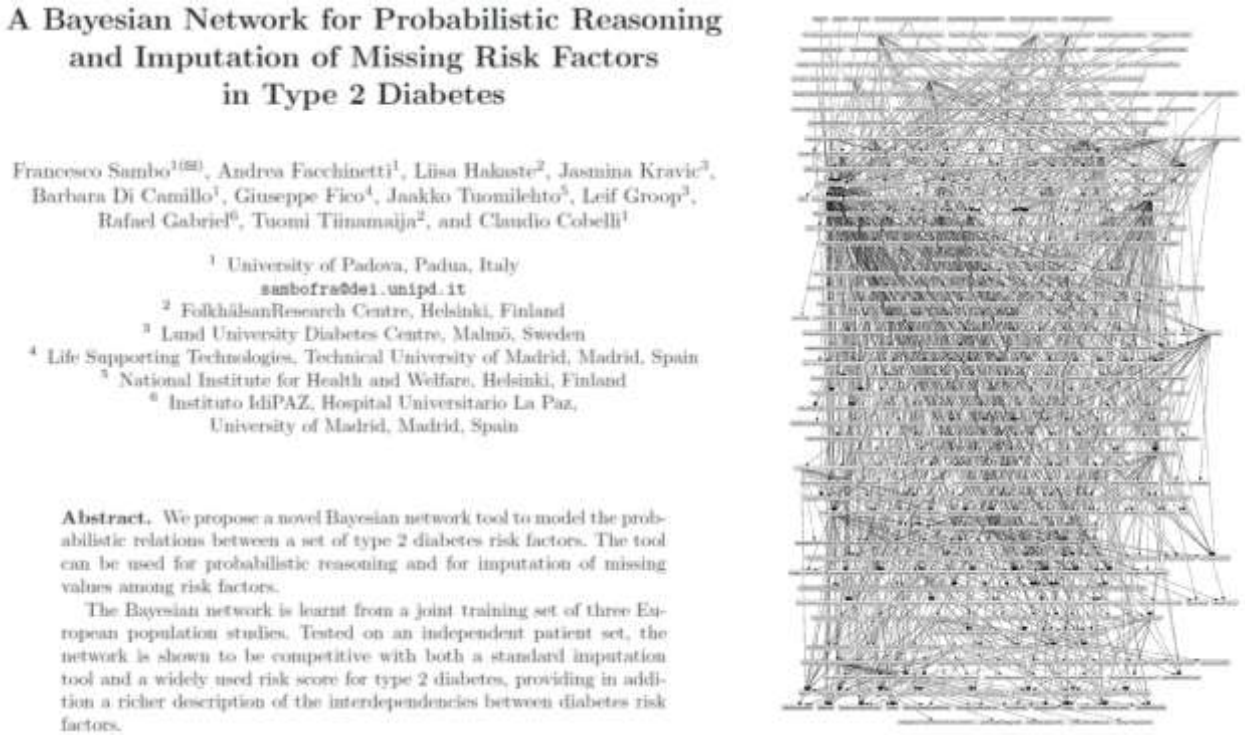
[70] La méthode relève toujours de l'élagage d'arbre de décisions dans les options de jeu avec un réseau de neurones qui s'améliore par renforcement en jouant contre lui-même. Voir l'article technique Mastering the Game of Go without Human Knowledge (42 pages) qui documente la prouesse et AlphaGo Zero: Learning from scratch, de DeepMind, qui vulgarise la performance. J'ai décortiqué AlphaGo Zero dans Les conséquences pratiques d'AlphaGo Zero en novembre 2017.

[71] Source de l'image : https://xkcd.com/399/.

[72] Comme le rappelle **Brian Bannon** en 2009 dans Unreasonable Effectiveness of Data.

La plupart des études scientifiques dans le domaine de la biologie et de la santé génèrent des corpus sous forme de résultats statistiques comme des gaussiennes d'efficacité de nouveaux médicaments. L'exploitation de la masse de ces résultats relève aussi d'approches bayésiennes.

Le cerveau met d'ailleurs en œuvre une logique bayésienne pour ses propres prises de décision, notamment motrices, les centres associés étant situés dans le cervelet tandis que le cortex cérébral gère la mémoire et les actions explicites, y compris le déclenchement de mouvements[73].



A Bayesian Network for Probabilistic Reasoning and Imputation of Missing Risk Factors in Type 2 Diabetes

Francesco Sambo[1(✉)], Andrea Facchinetti[1], Liisa Hakaste[2], Jasmina Kravic[3], Barbara Di Camillo[1], Giuseppe Fico[4], Jaakko Tuomilehto[5], Leif Groop[3], Rafael Gabriel[6], Tuomi Tiinamaija[2], and Claudio Cobelli[1]

[1] University of Padova, Padua, Italy
sambofra@dei.unipd.it
[2] FolkhälsanResearch Centre, Helsinki, Finland
[3] Lund University Diabetes Centre, Malmö, Sweden
[4] Life Supporting Technologies, Technical University of Madrid, Madrid, Spain
[5] National Institute for Health and Welfare, Helsinki, Finland
[6] Instituto IdiPAZ, Hospital Universitario La Paz, University of Madrid, Madrid, Spain

Abstract. We propose a novel Bayesian network tool to model the probabilistic relations between a set of type 2 diabetes risk factors. The tool can be used for probabilistic reasoning and for imputation of missing values among risk factors.

The Bayesian network is learnt from a joint training set of three European population studies. Tested on an independent patient set, the network is shown to be competitive with both a standard imputation tool and a widely used risk score for type 2 diabetes, providing in addition a richer description of the interdependencies between diabetes risk factors.

Les méthodes statistiques se sont fondues avec le temps dans les techniques du machine learning et du deep learning. Ces dernières reposent en effet toutes sur des modèles probabilistes pour identifier des objets, faire des prévisions ou des recommandations[74]. S'appuyant sur de la logique formelle, l'IA symbolique ne relève pas des probabilités. Mais des combinaisons sont possibles comme lorsque l'on intègre la notion de logique floue aux moteurs de règles.

Raisonnement automatique

Le raisonnement automatique fait partie du vaste champ de l'IA symbolique appliquant de la logique formelle. Cela faisait historiquement partie du champ de la **GOFAI**, pour "Good old fashion AI ». Elle s'opposait à l'approche connexionniste qui exploite le biomimétisme et les réseaux de neurones dans une approche probabiliste. L'approche symbolique appliquée au raisonnement automatique est plus rigoureuse mais difficile à mettre en œuvre et à généraliser car il est très difficile de collecter les règles d'un métier ou d'un domaine donné.

La formalisation du raisonnement humain remonte à **Aristote** et à l'identification de règles formelles utilisées dans l'argumentation philosophique, à base de syllogisme associant deux prémisses et une déduction (si A et B sont vrais alors C est vrai).

---

[73] Voir Le cerveau statisticien : la révolution bayésienne en sciences cognitives, de Stanislas Dehaene, cours du Collège de France (31 slides). Source de l'illustration : A Bayesian Network for Probabilistic Reasoning and Imputation of Missing Risk Factors in Type 2 Diabetes 2015 (2 pages).

[74] Voir Traditional statistical methods often out-perform machine learning methods for time-series forecasts de Paul Cuckoo, juillet 2018.

D'autres ont tenté de formaliser la notion même de raisonnement et de création de nouveaux concepts comme le théologien **Ramon Llull** en 1308 avec sa roue d'assemblage de concepts de base puis **Gottfried Leibniz** en 1666 avec son art de la combinaison[75].

Suivirent les travaux de **Georges Boole** au 19e siècle et son algèbre formalisant l'usage de règles de raisonnement, puis de nombreux développements théoriques, notamment autour de la logique formelle, des calculs de prédicats, de la logique du premier et du second ordre[76].

À vrai dire, on ne sait toujours pas dire si le raisonnement humain s'appuie sur une manipulation de symboles à haut niveau ou par assemblage de connexions de bas niveau dans le cerveau, que l'on appelle approche connexionniste ou sous-symbolique.

Les outils du raisonnement automatique en IA sont très divers avec les moteurs de règles, les arbres d'exploration et autres graphes, les ontologies et les outils de modélisation des connaissances. Pendant longtemps, les approches symboliques et connexionnistes se sont opposées.

Les recherches les plus récentes en IA visent à les rapprocher et d'au moins deux manières. Tout d'abord en exploitant le deep learning pour extraire des règles de documents en texte libre ou en associant textes et images dans des logiques multimodales, qui alimentent ensuite des moteurs de règles traditionnels. Et puis en cherchant à intégrer des briques de raisonnement dans le deep learning, le raisonnement étant l'aboutissement de techniques (toujours probabilistes) de traitement du langage et de gestion des connaissances.

Démonstrations de théorèmes

Les débuts des moteurs de règles à la base des systèmes experts remontent à 1957 quand **Alan Newell** et **Herbert Simon** développaient le General Problem Solver (GPS), un logiciel de résolution de problèmes mathématiques utilisant des règles modélisant les inférences possibles d'un domaine et résolvant un problème en partant de la solution attendue et en remontant vers les hypothèses. Ce GPS faisait suite au **Logic Theorist** produit début 1956 par ces mêmes Allen Newell, Herbert Simon et Cliff Shaw, quelques mois avant la création d'appellation d'intelligence artificielle par Jim MacCarthy. Le Logic Theorist démontrait quelques théorèmes mathématiques à base de logique axiomatique[77].

---

[75] Voir 12 Artificial Intelligence (AI) Milestones: 2. Ramon Llull And His 'Thinking Machine' par Gil Press, février 2020 qui dresse une intéressante histoire philosophique de la structuration du savoir.

[76] On distingue habituellement trois niveaux de logique : la logique d'ordre zéro ou logique des propositions, la logique d'ordre un ou calcul/logique des prédicats et la logique d'ordre deux. La logique d'ordre zéro exploite des propositions qui sont vraies ou fausses et sans variables. La logique d'ordre un raisonne à partir de variables et de quantificateurs. La logique d'ordre deux peut utiliser des fonctions. Tout ceci est à peu près correctement expliqué dans Une nouvelle logique mathématique ainsi que dans Intelligence Artificielle Symbolique de Guillaume Piolle, 2015 (109 slides). Mais cela reste assez théorique et difficile à appréhender.

[77] Voir le compte-rendu des auteurs sur Logic Theorist : The Logic Theory Machine A Complex Information Processing System, juin 1956 (40 pages).

En 1958, **John McCarthy** créait le langage de programmation LISP adapté en particulier à la création de moteurs de règles. Il a abouti à la création d'une petite industrie dans les années 1980 avec les ordinateurs spécialisés de **Lisp Machines** et **Symbolics** (1979-2005) et les logiciels d'**Intellicorp**[78].

Nous l'avons déjà cité, en 1959, **Herbert Gelernter** créait le Geometry Theorem Prover[79], capable de démontrer des théorèmes de géométrie et tournant sur un IBM 704. Le même Gelernter est à l'origine du **SYNCHEM** (SYNthetic CHEMistry), un système expert des années 1970 capable de déterminer des réactions chimiques de synthèse de molécules organiques.

Dans les années 1970, **Robert Kowalski** de l'Université d'Edimbourg fit avancer les travaux dans la programmation logique. Puis les français **Alain Colmerauer** et **Philippe Roussel** créèrent le langage de programmation **Prolog** (**Pro**grammation en **log**ique) qui connut ses heures de gloire pendant les années 1980 et 1990, et surtout dans l'éducation et la recherche. Prolog était principalement utilisé en Europe tandis que LISP l'était de manière dominante aux USA. Un *replay* de la bataille NTSC vs PAL/Secam dans la TV couleur !

```
mother_child(trude, sally).

father_child(tom, sally).
father_child(tom, erica).
father_child(mike, tom).

sibling(X, Y)      :- parent_child(Z, X), parent_child(Z, Y).

parent_child(X, Y) :- father_child(X, Y).
parent_child(X, Y) :- mother_child(X, Y).
```

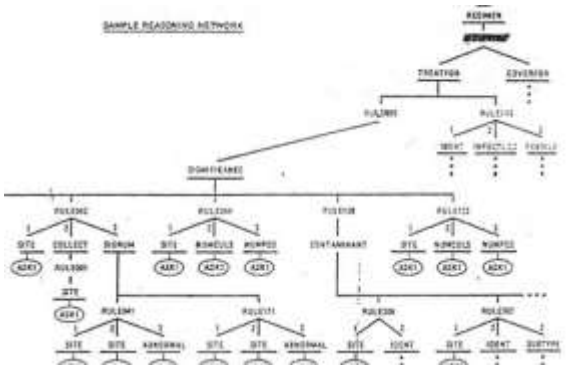
This results in the following query being evaluated as true:

```
?- sibling(sally, erica).
Yes
```

Prolog est un langage déclaratif qui sert à gérer de la logique du premier ordre avec la déclaration de relations, faits et règles. Il est notamment utilisé pour l'analyse du langage. Prolog est d'ailleurs utilisé à cet effet dans DeepQA, une des briques d'IBM Watson.

Premiers systèmes experts

Quelques expérimentations ont marqué les débuts des systèmes experts comme **MYCIN**, un système permettant de déterminer les bactéries responsables d'infections en fonction des symptômes (*ci-dessous*) avec une base de 450 règles.



```
RULE036
PREMISE: ($AND (SAME CNTXT GRAM GRAMNEG)
               (SAME CNTXTM MORPH ROD)
               (SAME CNTXT AIR ANAEROBIC))
ACTION: (CONCLUDE CNTXT IDENTITY BACTEROIDES TALLY 0.6)

IF: 1) The gram stain of the organism is gramneg, and
    2) The morphology of the organism is rod, and
    3) The aerobicity of the organism is anaerobic

THEN: There is suggestive evidence (0.6) that the identity
          of the organism is bacteroides
```

Les systèmes experts ont été théorisés dans le cadre du **Stanford Heuristic Programming Project** en 1980. Ils répondent à des questions dans des domaines spécifiques dont on a codifié la connaissance. Cela permit à l'IA de se rendre utile dans des domaines spécifiques, comme dans la santé ou la maintenance dans l'industrie[80].

---

[78] Créé en 1980 et maintenant spécialisé dans les logiciels de gestion d'applications pour SAP, un métier plus terre à terre

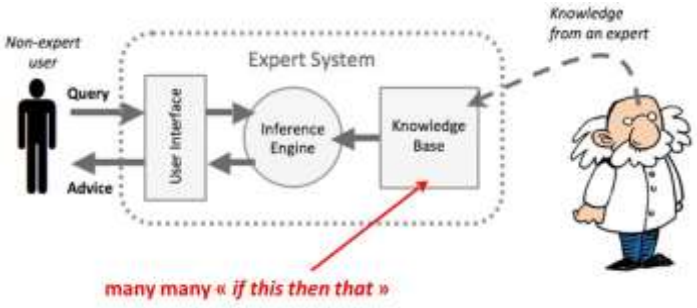
[79] Voir Realization of a Geometry-Theorem Proving Machine par Herbert Gelernter, 1959 (30 pages).

[80] Comme le système expert français SACHEM (Système d'Aide à la Conduite des Hauts fourneaux En Marche) chez Arcelor-Mittal qui pilote les hauts-fourneaux en exploitant les données issues de capteurs. Le projet a été mené entre 1991 et 1998 et serait toujours en opération.

Les moteurs de règles et les solveurs sont couramment employés dans les systèmes experts depuis les années 1980[81]. Et ils ont connu de nombreux progrès (*ci-dessous*) malgré l'hiver de l'IA de la fin des années 1980 et débuts 1990. C'était surtout un hiver des systèmes experts et du LISP !

DENDRAL – composition de matériaux, Feigenbaum, Buchanan et al, 1965
SYNCHEM – chimie organique, 1966-1997
ALICE – solver, Jean-Louis Laurière, 1976
PROSPECTOR – prospection géologique, 1977
R1 – Digital Equipment, 1982
DIPMETER – prospection géologique, 1982
MYCIN – diagnostique de maladies infectieuses, 1983
SOAR - Laird, 1983
Cyc – SI généraliste, Lenat and Guha, 1984
Neuron Data – Nxpert, 1985
SNARK4 – système expert, Jean-Louis Laurière, 1986
ILOG Rules - IBM Operational Decision Manager, 1987
EPIC - trafic aérien, Rosbe, Chong, and Kieras, 2001
Web sémantique / RDF - 2001
ACT-R - Anderson and Lebiere, 2003
ICARUS - Langley, 2005
SNePS - Semantic Network Processing System, Shapiro, 2007

Les moteurs de règles s'appuient sur la notion de raisonnement contraint par des règles et exploitant des bases de faits. On fournit au moteur un ensemble de règles et de faits pouvant par exemple représenter le savoir des experts dans un domaine donné. Avec des règles proches de la programmation logique du genre "*si X et Y sont vrais, alors Z est vrai*" ou "*X entraîne Y*".



On peut alors interroger le système en lui posant des questions genre "*est-ce que W est vrai ?*" et il va se débrouiller pour exploiter les règles enregistrées pour répondre à la question. Les moteurs de règles utilisent la théorie des graphes et la gestion de contraintes.

Un système expert s'appuie sur deux composantes clés : une **base de connaissance**, générée souvent manuellement ou éventuellement par exploitation de bases de connaissances existantes, et un **moteur de règles**, plus ou moins générique, qui va utiliser la base de connaissance pour répondre à des questions précises.

Les systèmes experts peuvent expliquer le rationnel de leur réponse. La traçabilité est possible jusqu'au savoir codifié dans la base de connaissances, un avantage que les réseaux de neurones du deep learning et le machine learning n'ont pas encore.

Systèmes experts d'aujourd'hui

On compte encore des outils et langages dans ce domaine et notamment l'offre du français **ILOG**, acquis en 2009 par IBM et dont les laboratoires de R&D sont toujours à Gentilly près de Paris, au sud du boulevard Périphérique. Le moteur d'inférence ILOG JRules est devenu **IBM Operational Decision Manager**. De son côté, ILOG Solver est une bibliothèque C++ de programmation par contraintes, devenue **IBM ILOG CPLEX CP Optimizer**. Une stratégie de branding moins efficace que celle d'IBM Watson, comme nous le verrons bien plus loin.

La mise en place de systèmes experts se heurtait à la difficulté de capter la connaissance des experts. Les temps de calcul pour les faire fonctionner étaient également longs avec les ordinateurs de l'époque. La loi de Moore a permis de limiter ce dernier écueil depuis.

[81] On peut citer notamment l'outil de développement Nexpert pour Macintosh et PC de **Neuron Data**, une startup créée en 1985 aux USA par les français Alain Rappaport, Patrick Perez et Jean-Marie Chauvet. Elle a été revendue au début des années 2000à l'Allemand Brokat puis à différents acquéreurs successifs avant de disparaître.

Il existe d'autres types de systèmes experts qui mettent en œuvre la notion de programmation par contraintes, permettant d'atteindre un objectif en fonction d'une base de règles, d'objectifs et de contraintes opérationnelles. Dans de nombreux domaines, la force brute et le deep learning se sont ensuite imposés en lieu et place de la logique formelle et de la captation manuelle de connaissances.

moteurs de règles

open source

CLIPS : moteur de règles dans le domaine public.

Drools : distribué par Red Hat.

DTRules : moteur de règles en Java.

Gandalf : moteur de règles tournant sur PHP.

OpenL Tablets : business centric rules and BRMS.

propriétaires

Corticon : moteur de règles sous Java et .NET, filiale de Progress Software.

IBM Operational Decision Manager : ex ILOG Rules.

JESS : moteur de règle Java, sur-ensemble du langage CLIPS.

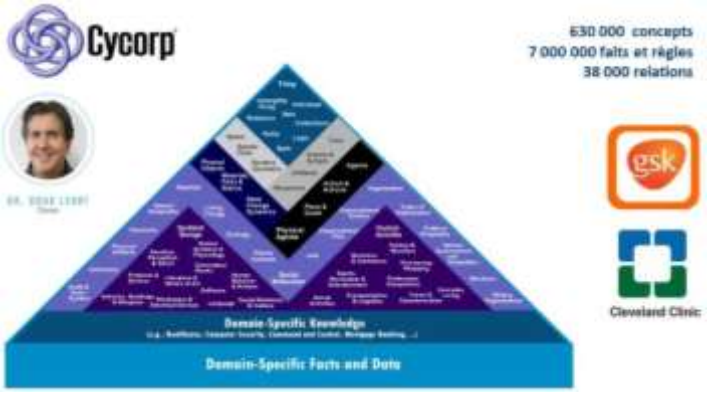
Microsoft Azure Business Rules Engine : framework de moteur de règle en .NET.

Oracle Policy Automation : modélisation et déploiement de règles.

Les logiciels de moteurs de règles du marché sont appelé BRMS pour **Business Rules Management Systems**. L'offre est assez abondante mais plus ancienne et moins connue que celle qui concerne le machine learning et le deep learning (*ci-dessous*). Cette offre de BRMS est maintenant intégrée dans le concept plus large de **Decision Management Systems** qui associent des moteurs de règles et des outils d'*analytics*.

L'un des systèmes experts les plus ambitieux des années 1980 était **Cyc**.

Il devait comprendre une énorme base de connaissances de centaines de milliers de règles. Ce projet était piloté par Doug Lenat du consortium de recherche privé MCC qui ferma ses portes en 2000. Doug Lenat l'a transformé en projet entrepreneurial avec **Cycorp**, lancée en 1994[82]. Cette dernière propose une base de connaissance intégrant 630 000 concepts, 7 millions de faits et règles et 38 000 relations.



Le tout étant exploitable par des moteurs de règles. La base est notamment alimentée par l'analyse de documents disponibles sur Internet. Mais ce projet est considéré comme un échec.

Cycorp est une sorte de laboratoire de recherche privé en IA financé par des contrats du gouvernement US, dont la DARPA, et d'autres pour des entreprises privées. Il propose une suite d'outils en open source et licence commerciale permettant d'exploiter des dictionnaires, ontologies et bases de connaissances pour répondre à des questions d'analystes. Le système expert OpenCyc 4.0 qui exploitait la base de Cycorp n'est plus disponible en open source depuis 2017 pour éviter le fameux phénomène de la fragmentation du code (« fork »). Il est depuis commercialisé sous forme de licences dédiées à la recherche ou de licences commerciales. Cycorp est devenu une société de conseil et d'intégration en IA, spécialisée notamment dans le traitement du langage et des connaissances. Bref, pour l'instant, cela ne va pas fort pour eux.

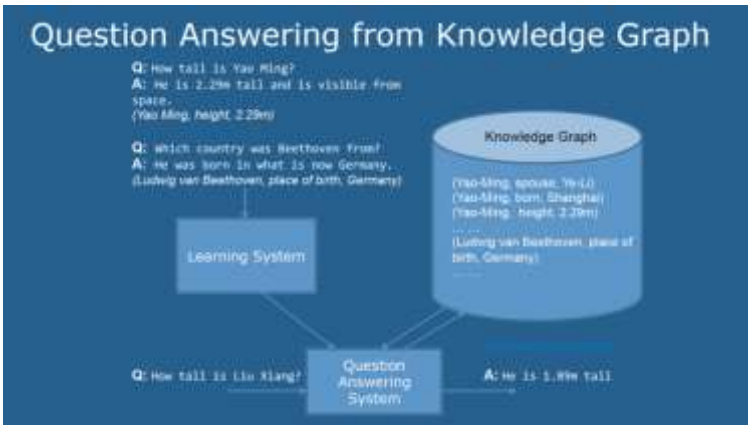
L'initiative open source **Schema.org** lancée par Google, Microsoft, Yahoo et le Russe Yandex propose de son côté des millions de types, descriptions de faits exploitables par les moteurs de recherche et les moteurs de règles. Il permet notamment aux web masters de standardiser la nomenclature utilisée dans les descriptifs de pages web.

Les outils dotés de capacités de raisonnement continuent d'évoluer pour faire avancer le champ de la représentation des connaissances et du raisonnement.

[82] Voir ce talk à TEDx Youth Austin de Doug Lenat qui date de 2015 (16 mn) et The 'Cyc' project and why it was flawed par W Daniel Hillis qui explique pourquoi la démarche de Doug Lenat ne pouvait pas aboutir (2 mn).

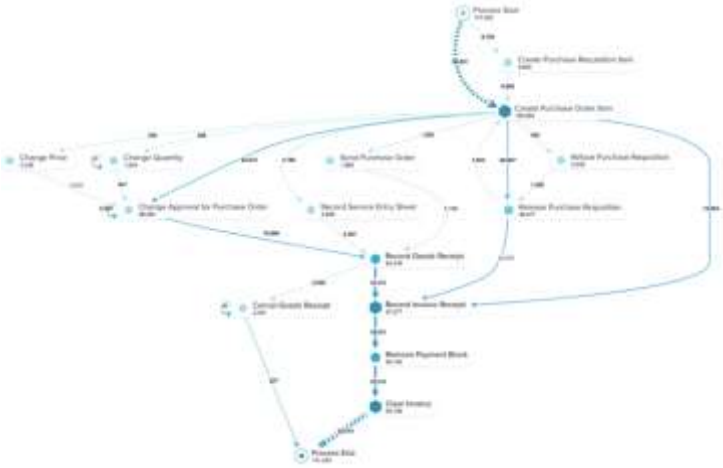
Les techniques associées sont moins connues que celles du machine learning et du deep learning, ce d'autant plus qu'elles sont de plus en plus hybrides. Ainsi, un moteur de règles peut-il exploiter des règles elles-mêmes générées par analyse du langage dans des réseaux de neurones récurrents.

Le deep learning et les réseaux de neurones récurrents que nous verrons plus loin alimentent maintenant couramment les bases de connaissances et les moteurs de règles qu'ils ont contribué indirectement à faire décliner. Les bases de connaissances sont devenues des graphes de connaissances (**Knowledge Graphs)**, mais c'est un peu la même chose, comme l'illustre le schéma *ci-contre*[83], issu d'une présentation sur les réseaux de neurones symboliques.



Faisons le tour de quelques sociétés du secteur.

Celonis (2011, Allemagne, $77,5M) propose une solution de « process mining » qui sert à reconstruire les processus de workflow d'une entreprise comme celui du règlement fournisseurs, du service clients ou de flux de travail collaboratif, le tout par analyse de logs. Le processus est ensuite analysable avec détection de variations et anomalies via des algorithmes de machine learning, puis donne lieu à des optimisations. Bref, de la donnée brute permet de générer des graphes de workflows.



Khresterion (2014, France) propose un logiciel d'aide au diagnostic dans divers domaines. Ils s'appuient sur un moteur de règles, K Engine, qui exploite une représentation de la connaissance sous forme d'ontologies et adopte une structure en graphe et non d'arborescence, ce qui la rend très ouverte. Cela permet de gérer la contradiction et la non complétude d'informations. La société travaille dans les domaines financiers et juridiques après avoir tâté du domaine de la santé.

ExpertSystem (1989, Italie) propose diverses solutions pour les entreprises à base de systèmes experts et d'outils de traitement du langage naturel.

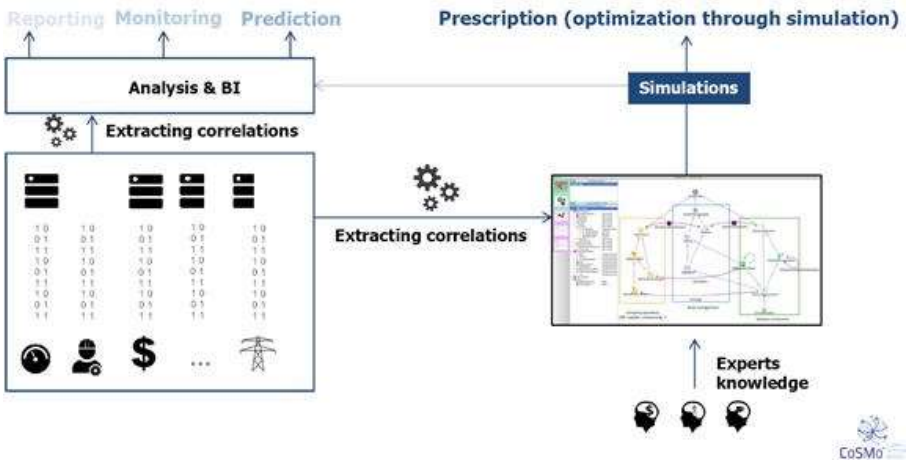
Tree-Logic (1986, France) propose « la maïeutique », une solution de raisonnement automatique crée par Jean-Philippe de Lespinay qui se débat depuis plus de trois décennies pour faire adopter ses vues. Le principe de sa solution repose sur l'extraction de connaissances implicites des experts et d'usage d'une logique d'ordre zéro. La solution est mise en forme dans un agent conversationnel qui exploite un système expert de quelques centaines de règles métier.

[83] Ce schéma est issu de Toward Neural Symbolic Processing de Hang Li, 2017 (36 slides).

Elle avait été déployée il y a longtemps à la Banque de Bretagne et chez quelques autres[84].

Cosmo Tech (2010, France/USA, $32,9M) est une startup spin-off de l'ENS Lyon et du CNRS basée à Lyon et aux USA qui a développé une plateforme logicielle de modélisation et de simulation de systèmes complexes. Elle s'appuie sur le langage de modélisation CosML qui sert à représenter les états ainsi que les comportements des systèmes complexes et à les étudier grâce à de la simulation. Le système exploite des règles métier et des corrélations extraites de données de production via des techniques de machine learning (*schéma ci-dessous*).

La solution est déclinée dans diverses industries comme avec leur application *Asset Investment Optimization* (AIO) dédiée aux énergéticiens, *Crisis Management* qui permet la gestion de crise et *Smart Territories* qui permet de modéliser des systèmes complexes pour la ville intelligente.



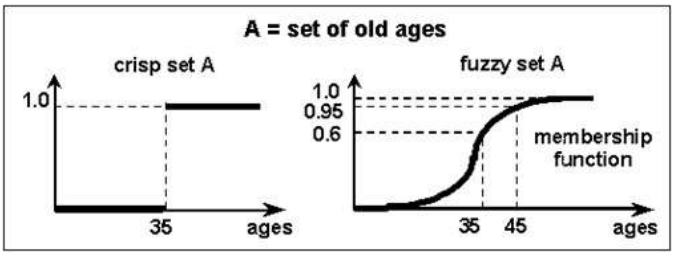
C'est un excellent exemple d'hybridation technologique illustrant la manière dont les systèmes experts s'intègrent dans les solutions d'IA.

Comme nous le verrons dans la partie dédiée à l'AGI, des tentatives nombreuses visent à utiliser des réseaux de neurones pour faire du raisonnement symbolique. L'un des premiers du genre fut le **KBANN** (Knowledge-based Artificial Neural Network) de Towell et Shavlik en 1994[85] mais qui ne semble pas avoir abouti à des applications pratiques.

Logique floue

La logique floue est un concept de logique inventé par l'américain **Lofti Zadeh** ("Fuzzy Logic") en 1965[86].

Elle permet de manipuler des informations floues qui ne sont ni vraies ni fausses, en complément de la logique booléenne, mais à pouvoir faire des opérations dessus comme l'inversion, le minimum ou le maximum de deux valeurs ou le regroupement d'ensembles. On peut aussi faire des OU et des ET sur des valeurs "floues".

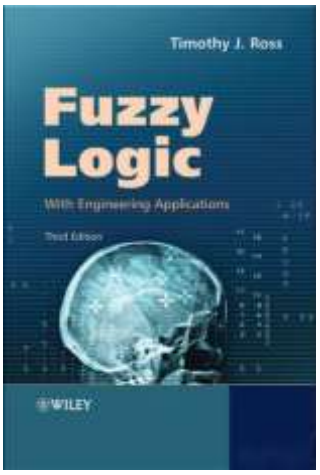


[84] Voir Notre technologie et L'ordinateur intelligent, une invention française bloquée depuis 20 ans, par Bernard Lambilly, Les Echos, 2011, qui raconte l'histoire fort chahutée de cette invention. Jean-Philippe de Lespinay se débat contre l'establishment de la recherche en IA qui mépriserait ses travaux, comme en témoigne sa difficulté à faire valider une fiche le concernant dans la version française de Wikipedia. Sa fiche Wikipedia en anglais dont il est visiblement l'auteur a été supprimée car s'appuyant essentiellement sur des auto-références. L'auteur décrit sa méthode dans un article publié dans Science & Vie en 1991, Intelligence artificielle : du zéro pointé au « zéro plus ». L'auteur m'a harcellé plusieurs fois, clamant haut et fort que son IA française était la seule véritable IA du marché, à l'exclusion donc de toutes les autres formes et techniques d'IA.

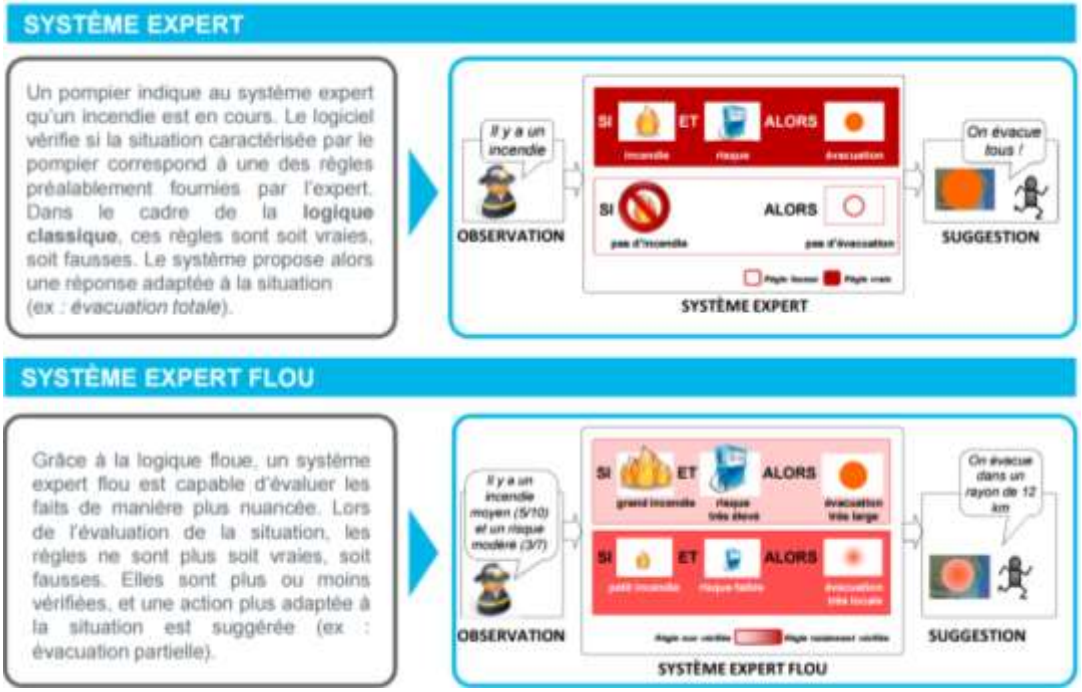
[85] Voir Knowledge-Based Artificial Neural Networks, 1994 (45 pages) qui est cité dans Reasoning with Deep Learning: an Open Challenge de Marco Lippi, 2016 (22 slides).

[86] J'avais eu l'occasion de l'entendre la présenter lors d'une conférence à l'Ecole Centrale en 1984, lorsque j'étais en option informatique en troisième année. Ça ne nous rajeunit pas ! Lotfi Zadeh est décédé en septembre 2017. Voir Lotfi Zadeh 1921–2017, par Richard Lipton et Ken Regan, octobre 2017.

Quid des applications de la logique floue[87]? On les trouve dans le contrôle industriel[88], dans des boîtes de vitesse chez **Volkswagen** (pour tenir compte de l'intention "floue" du conducteur), pour gérer des **feux de circulation** et maximiser le débit, dans de nombreuses applications du BTP, dans la reconnaissance de la parole et d'images, le plus souvent, en complément du bayésien et dans des outils de recommandation. Des dizaines de milliers de brevets auraient été déposés pour protéger des procédés techniques utilisant la théorie de la logique floue.



Les moteurs de règles de systèmes experts peuvent d'ailleurs intégrer les principes de la logique floue (*ci-dessous*). Ceci dit, la logique floue n'est pas utilisée de manière très courante, notamment du fait que les systèmes experts ne sont plus à la mode depuis une quinzaine d'années.



Lotfi Zadeh
1921-2017

J'ai identifié la startup **Zsolusionz** qui proposait des solutions à base de logique floue mais qui n'a plus de traces sur Internet. On peut compter également sur **intellitec** (1998, France), une société de R&D spécialisée dans la logique floue et ses applications industrielles, fondée par Zyed Zalila qui enseigne à l'UTC Compiègne. La société avait démarré en créant des systèmes d'aide à la conduite (ADAS). Leur offre s'articule autour du moteur de règles floues xtractis. L'un de ses bénéfices est de créer des systèmes d'IA explicables dotés de capacités d'apprentissage inductif automatique[89].

Recherche opérationnelle

La recherche opérationnelle est une autre branche de l'IA symbolique. Elle vise à résoudre des problèmes de logique et d'optimisation qui sont très nombreux dans les entreprises, notamment dans les utilities et les transports.

---

[87] Quelques exemples d'applications de la logique floue dans cette présentation : Applications of fuzzy logic de Viraj Patel, 2016 (22 slides).

[88] Voir cette référence : Fuzzy Logic with Engineering Applications par Timothy J. Ross, 2010 (607 pages).

[89] Voir le livre blanc Applications opérationnelles des mathématiques du flou par Zyed Zalila, 2018 (11 pages) qui fait un inventaire d'études de cas de la logique flou dans une vingtaine de marchés. Ainsi que Approche xtractif par Zyed Zalila et al, 2018 (18 pages).

Comment allouer ses ressources pour qu'elles soient les mieux utilisées ? Comment maximiser son revenu avec ses ressources existantes ? Comment optimiser le parcours d'un livreur ou d'un commercial ?

Les solutions pour résoudre ces problèmes sont variées. Elles peuvent faire appel à des algorithmes ad-hoc, à de la simulation, à des réseaux d'agents intelligents, une méthode qui est à la frontière entre l'intelligence artificielle et l'étude des systèmes complexes[90]. Il existe des outils de résolution de problèmes de logique comme les problèmes SAT[91].

Là encore, on se retrouve rapidement dans des situations d'hybridation d'algorithmes et technologies pour résoudre des problèmes qui intègrent des règles de logique et des contraintes numériques (*aka* : programmation par contrainte, logique propositionnelle, logiques monotone et non-monotone, des notions de combinatoires et d'exploration d'arbres de décision), le traitement du langage et la modélisation des connaissances.

Machine learning

Le vaste domaine du machine learning, ou apprentissage automatique, ou encore apprentissage statistique, est défini comme le champ de l'IA qui utilise des méthodes probabilistes pour apprendre à partir des données.

D'un point de vue pratique, le machine learning vise notamment à reconnaître et classifier des objets (des lettres, des objets dans des images), à faire des prévisions sur des données (régressions linéaires et non linéaires), à découvrir des corrélations entre données et événements (comme pour déterminer si un logiciel est un virus, si un client risque de quitter un service sur abonnement ou au contraire, s'il sera intéressé par telle ou telle offre ou qu'un tableau clinique d'un patient est symptomatique de l'émergence d'une pathologie de longue durée), à segmenter des jeux de données (comme une base clients), le tout en exploitant des données d'entraînement.

Le machine learning est utilisé lorsque l'on ne sait pas modéliser un système avec des équations déterministes permettant d'obtenir directement les résultats recherchés. Cela s'applique donc surtout aux systèmes complexes qui ne sont ni modélisables ni simulables.

Le père fondateur du machine learning est **Arthur Samuel,** qui en définit le terme en 1959. Le machine learning donne aux machines la capacité d'apprendre sans être explicitement programmées. Le machine learning requiert presque toujours de faire des choix de méthodes et des arbitrages manuels pour les data scientists et les développeurs de solutions. Le choix des méthodes reste pour l'instant manuel, même si certaines startups proposent maintenant d'automatiser ce processus. Et on continue de coder pour créer des solutions de machine learning !

L'apprentissage automatique s'appuie sur des données existantes. Elles lui permettent de produire des prévisions, des segmentations ou des labels à partir de la généralisation d'observations. La qualité et la distribution statistique des données d'entraînement conditionnent celle des prévisions. Si les données ne représentent pas correctement l'espace du possible, les prévisions ou classifications ne seront pas bonnes et seront aussi biaisées. L'entraînement de modèles de machine learning consiste à réduire au minimum la **fonction de coût** (cost function) qui évalue l'erreur dans les opérations d'entraînement par apprentissage supervisé.

Les données d'entraînement sont donc absolument critiques pour la qualité des résultats. Un bon système de machine learning doit pouvoir s'adapter à différentes contraintes comme une évolution permanente des données d'entraînement, ainsi que leur incomplétude et leur imperfection.

---

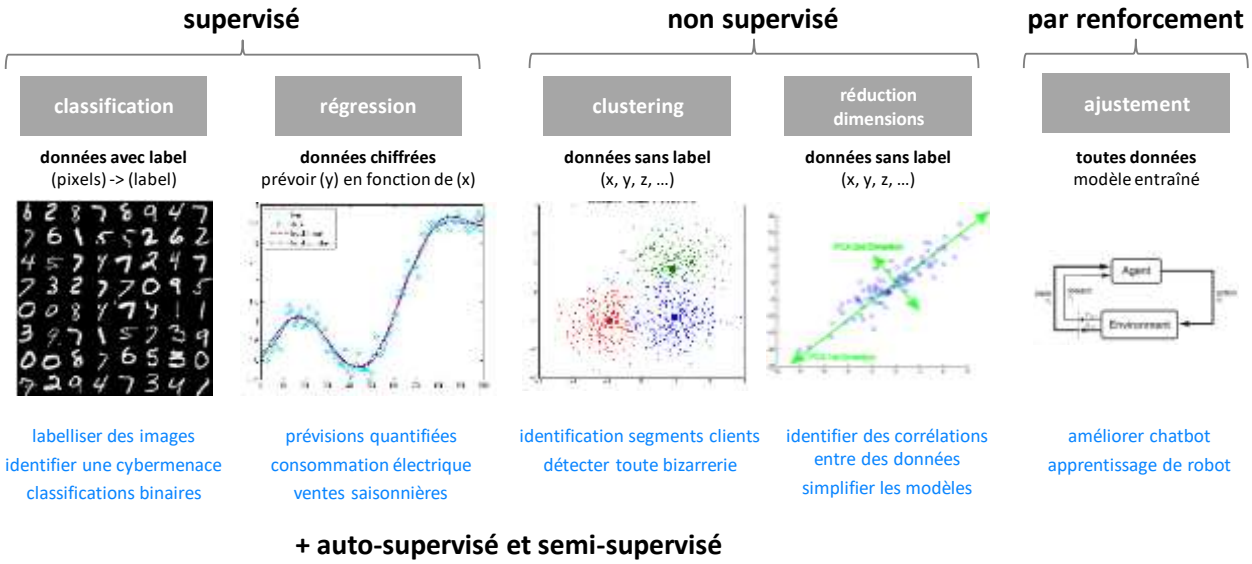
[90] Source : Renouveau de l'intelligence artificielle et de l'apprentissage automatique, un rapport de l'Académie des Technologies coordonné par Yves Caseaux et publié en avril 2018.

[91] Voir Ils ne savaient pas que c'était insoluble, alors ils l'ont résolu, par Serge Abiteboul et Charlotte Truchet, octobre 2019.

On distingue en général trois grandes catégories de machine learning selon les méthodes d'apprentissage utilisées, mais qui conditionnent aussi leur fonctionnalité, reprenant le schéma *ci-dessous* :

- L'**apprentissage supervisé** avec la classification qui permet de labelliser des objets comme des images et la régression qui permet de réaliser des prévisions sur des valeurs numériques. L'apprentissage est supervisé car il exploite des bases de données d'entraînement qui contiennent des labels ou des données contenant les réponses aux questions que l'on se pose. En gros, le système exploite des exemples et acquiert la capacité à les généraliser ensuite sur de nouvelles données de production. C'est la méthode la plus couramment utilisée. L'un des enjeux clés de ce domaine est de diminuer les besoins en données pour réaliser l'apprentissage, notamment dans le cas particulier du deep learning.
- L'**apprentissage non supervisé** avec le clustering et la réduction de dimensions. Il exploite des jeux de données non labellisées pour son entraînement. Ce n'est pas un équivalent fonctionnel de l'apprentissage supervisé qui serait automatique. Ses fonctions sont différentes. Le clustering permet d'isoler des segments de données spatialement séparés entre eux, mais sans que le système puisse les labelliser ou en fournir une explication sémantique. A charge pour les humains ou à du machine learning à apprentissage supervisé de déterminer ces labels. La réduction de dimensions vise à réduire la dimension de l'espace des données utilisées, en choisissant les dimensions les plus pertinentes. Du fait de l'arrivée de la « big data », la dimension des données a explosé et les recherches sur les techniques d'embedding associées sont très actives.
- L'**apprentissage par renforcement** pour l'ajustement de modèles déjà entraînés en fonction des réactions de l'environnement. C'est une forme d'apprentissage supervisé incrémental qui utilise des données arrivant au fil de l'eau servant à adapter le comportement du système. C'est utilisé par exemple en robotique, dans les jeux ou dans les chatbots capables de s'améliorer en fonction des réactions des utilisateurs. Et le plus souvent, avec le sous-ensemble du machine learning qu'est le deep learning. L'une des variantes de l'apprentissage par renforcement est l'apprentissage supervisé autonome notamment utilisé en robotique où l'IA entraîne son modèle en déclenchant d'elle-même un jeu d'actions pour vérifier ensuite leur résultat et ajuster son comportement[92].

Voici un schéma maison qui résume tout cela de manière visuelle :



[92] L'apprentissage par renforcement peut aussi imiter le fonctionnement du cerveau humain qui fonctionne sur un système de récompenses à base de l'hormone dopamine.Voir An algorithm that learns through rewards may show how our brain does too par Karen Hao, janvier 2020.

Nous allons explorer une par une ces différentes méthodes de machine learning et les modèles mathématiques sur lesquelles elles s'appuient. Nous couvrirons aussi les variantes comme les **apprentissages auto-supervisé et semi-supervisé** qui s'appuient généralement un mix d'apprentissage supervisé et d'apprentissage non supervisé avec une partie seulement des données d'entraînement qui sont labellisées.

Classification

Il s'agit de pouvoir associer une donnée complexe comme une image ou un profil d'utilisateur à une classe d'objet, les différentes classes possibles étant fournies a priori par le concepteur du système.

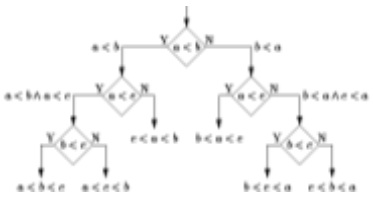
La classification utilise un jeu de données d'entraînement associé à des descriptifs (les noms des classes d'objets) pour la détermination d'un modèle. Cela génère un modèle de machine learning qui permet de prédire la classe d'une nouvelle donnée fournie en entrée. Un modèle est en fait un logiciel avec les paramètres qui lui permettent de fonctionner.

Dans les exemples classiques, nous avons la reconnaissance d'un simple chiffre dans une image, l'appartenance d'un client à un segment de clients ou pouvant faire partie d'une typologie particulière de clients (mécontents, pouvant se désabonner à un service, etc) ou la détection d'un virus en fonction du comportement ou de caractéristiques d'un logiciel. Un cas plus exotique peut consister à identifier des préférences politiques d'une personne en fonction de la photo de son véhicule, évidemment avec un taux d'erreur probablement significatif.

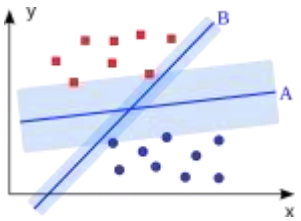
Comme nous le verrons plus loin, le machine learning se découpe lui-même en sous catégories : le machine learning de base sur des données avec un petit nombre de dimensions, que nous verrons ici, puis les réseaux de neurones et le deep learning, que nous traiterons plus loin, et qui sont adaptés aux données comportant un grand nombre de dimensions comme les images.

Il existe plusieurs méthodes de classification dont voici les principales.

Les **arbres de décision** que l'on appelle aussi les CART (Classification And Regression Tree) exploitent des critères discriminants, comme dans un moteur de règles. Ils permettent de classifier un objet en se posant successivement des questions (comparaison de données, …). Il en existe plusieurs sortes, telles que les CHAID (pour CHI-square Adjusted Interaction Detection) qui peuvent utiliser des branches multiples à chaque nœud.



arbres de décision



Support Vector Machines

$$P(c_j|d_i)=\frac{P(c_j)\prod_{k=1}^{|d_i|}P(w_{d_k}|c_j)}{\sum_{r=1}^{|C|}P(c_r)\prod_{k=1}^{|d_i|}P(w_{d_k}|c_r)}$$

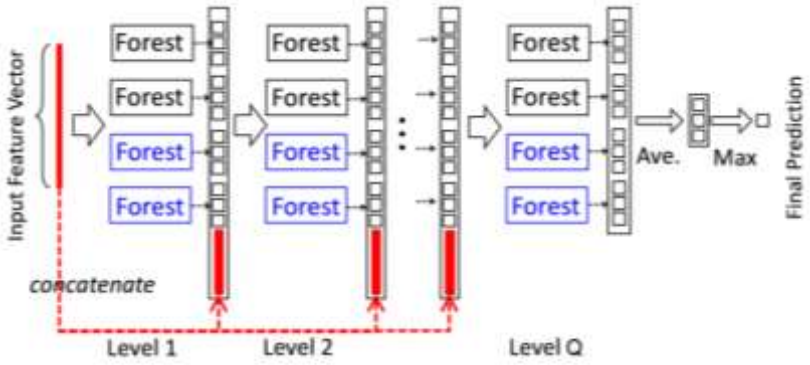
classification Bayésienne naïve

Ensemble Methods

Les arbres de décision sont déclinés dans des versions multiples avec notamment les **Random Forests** (qui utilisent plusieurs arbres de décision de classification d'objet basés sur l'usage de fonctions de classification aléatoires et font la moyenne des résultats, sorte de variation de la méthode des ensembles[93]) et le **Stacking** (qui utilise un empilement de plusieurs niveaux d'arbres de décision).

[93] Les random forests ont été proposés en 1995 par Tin Kam Ho puis étendus en 2001 par Leo Breiman de Berkeley et Adele Cutler de l'Université de l'Utah. C'est même une marque déposée par ces deux inventeurs. Voir Random Forests par Leo Breiman, 2001 (33 pages) et Trees and Random Forests par Adele Cutler, 2013 (92 slides).

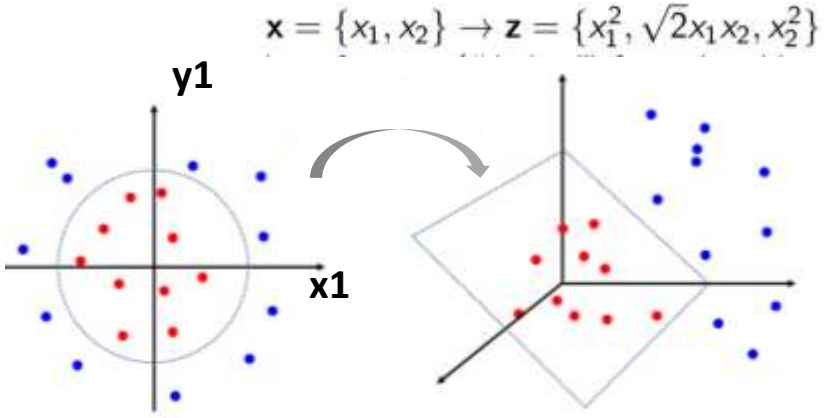
Le **Deep Forest** est la version la plus sophistiquée. **C'**est un empilement de couches d'ensembles de random forests, chacune servant à identifier des caractéristiques (features) qui ensuite sont transmises à la couche suivante, un peu comme dans les réseaux convolutifs. Cela peut s'appliquer à différents types de données.



Le Deep Forest permet un entraînement avec un plus faible volume de données et se positionne comme alternative au deep learning à base de couches de réseaux de neurones[94], que nous verrons plus loin. L'autre avantage est de permettre de créer des systèmes plus facilement explicables, du fait que les règles utilisées dans chaque arbre de décision sont en clair et pas le résultat d'un entraînement de réseaux de neurones par rétropropagation de gradients.

Les **Support Vector Machines** (SVM), créés en 1992 par Vladimir Vapnik, Isabelle Guyon et Bernhard Boser, servent à identifier une droite ou un hyperplan dans le cas d'un modèle à plusieurs dimensions qui permette de distinguer les classes d'objets les unes des autres de manière binaire en essayant de les séparer par une marge, représentée par une bande autour de l'hyperplan, aussi large que possible. On les appelle aussi des modèles d'indépendance conditionnelle simples.

Les SVM peuvent utiliser un modèle **non linéaire** lorsque les objets à séparer dans l'espace ne peuvent pas être isolés de part et d'autre d'un hyperplan. On recherche alors une fonction qui va transformer ces données, par exemple en 2D (x1, x2), dans un espace à deux ou trois dimensions, les dimensions étant un polynôme de x1 et x2, qui va permettre une séparation des données par un hyperplan si on ajoute une dimension[95].



Les **classifications naïves bayésiennes** utilisent les probabilités pour départager les objets dans des classes en fonction de caractéristiques bien établies[96]. À chaque hypothèse de départ, on associe une probabilité. L'observation d'une ou de plusieurs instances peut modifier cette probabilité. On peut parler de l'hypothèse la plus probable au vu des instances observées. Les probabilités bayésiennes présupposent l'indépendance des attributs utilisés.

---

[94] Voir Deep Forest: Towards An Alternative to Deep Neural Networks par Zhi-Hua Zhou et Ji Feng, 2017 (7 pages) et Deep Forest des mêmes auteurs, 2018 (34 pages), qui compare la performance de deep forest avec du deep learning, Deep Forest as a framework for a new class of machine-learning models, Lev Utkin & Al, 2018, The Deep Forest and its Modifications, de Lev Utkin, 2017 (66 slides) et A Report on Decision Tree, Random Forest and Deep Forest de Shrutina Agarwal.

[95] Source : Kernel Methods and Nonlinear Classification de Piyush Rai, un cours de Stanford, 2011 (95 slides) ainsi que SVM and Kernel machine - linear and non-linear classification de Stéphane Canu, 2014 (78 slides). L'exemple donné en schéma pourrait être traité avec une méthode plus simple consistant à transformer les coordonnées (x, y) en coordonnées polaires avec longueur vecteur + angle. On conserverait deux dimensions et la séparation liénaire SVM pourrait alors fonctionner.

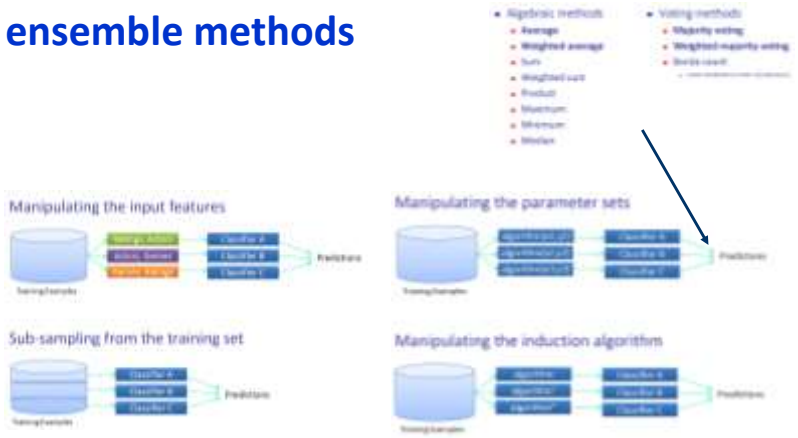
[96] La fiche Wikipedia de la classification naïve bayésienne explique bien son principe. Elle est inspirée des travaux de Thomas Bayes (1702 – 1761) repris ensuite par Laplace. Voir également la présentation Naïve Bayes Classifier (37 slides).

Elles appliquent le théorème de Bayes selon lequel Probabilité (A sous condition de B) = Probabilité (B sous condition de A)* Probabilité (A) / Probabilité (B). Ca fonctionne bien seulement si les probabilités de A et B sont bien indépendantes, ce qui n'est pas toujours le cas dans la vraie vie.

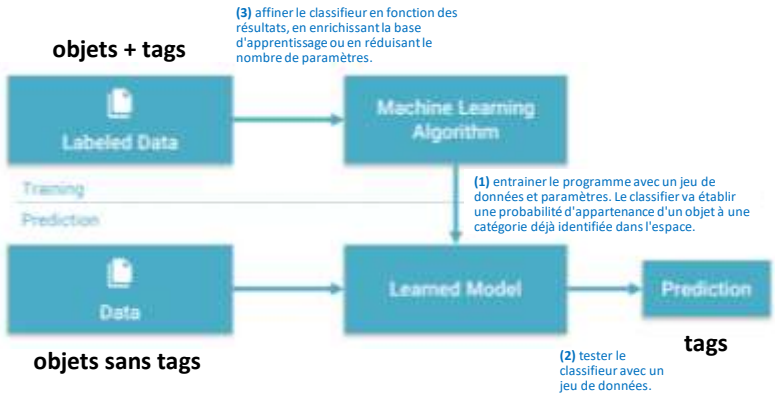
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Enfin, les **méthodes des ensembles** combinent plusieurs méthodes de classification pour en panacher les résultats et renforcer le poids des meilleures méthodes sans dépendre d'une seule d'entre elles.

Les méthodes des ensembles peuvent combiner des méthodes dites algébriques (avec une moyenne, une moyenne pondérée, un maximum, un minimum ou une médiane) et des méthodes par vote (utilisant la majorité, un vote pondéré, …), comme représenté dans le schéma *ci-contre*[97]. Le choix des assemblages dépend de la distribution statistique des données d'entraînement.



Un modèle mathématique de machine learning est entraîné avec un jeu de données d'apprentissage. Cet entraînement consiste à déterminer la bonne méthode à utiliser ainsi que les paramètres mathématiques du modèle retenu. Il va générer un modèle entraîné, et ses variables de fonctionnement pour faire une prévision ou une classification.



Le modèle entraîné est ensuite alimenté avec de nouveaux objets pour prédire leur appartenance à une classe déjà identifiée. Les spécialistes du machine learning testent habituellement différentes méthodes de classification pour identifier celle qui est la plus efficace compte-tenu du jeu de données d'entraînement, c'est-à-dire, celle qui génère un maximum de bonnes réponses pour un test réalisé avec un jeu de données en entrées qui sont déjà classées mais qui n'ont pas servi à l'entraînement du modèle.

Régression

Une régression permet de prédire une valeur numérique y en fonction d'une valeur x à partir d'un jeu d'entraînement constitué de paires de données (x, y). x peut être une seule valeur numérique ou une combinaison de valeurs (vecteur ou matrice de données). On peut par exemple prédire la valeur d'un bien immobilier ou d'une société en fonction de divers paramètres les décrivant.

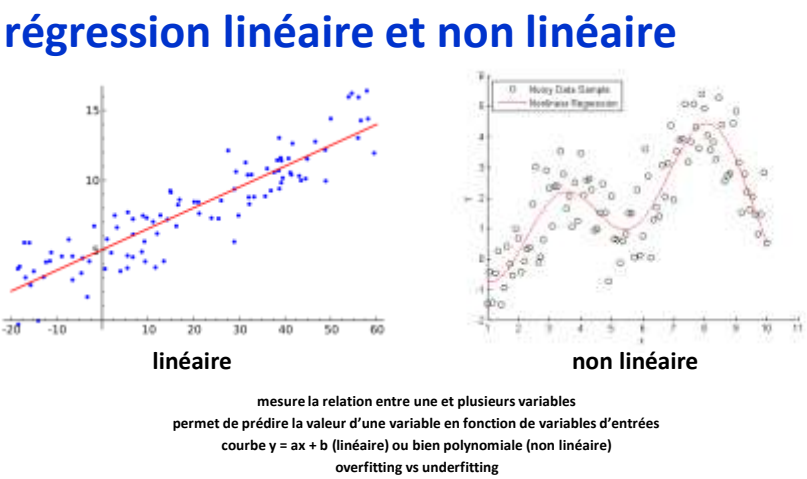
[97] En pratique, on utilise diverses méthodes d'agrégation de résultats avec les adaboost, boosting, gradient boosting, XGBoost, bagging, LightGBM de Microsoft et autres random forest. C'est bien expliqué dans Ensemble Learning to Improve Machine Learning Results de Vadim Smolyakov, août 2017.

Les schémas *ci-dessous* qui illustrent ce concept utilisent uniquement une donnée en entrée et une en sortie. Dans la pratique, les régressions utilisent plusieurs paramètres en entrée. Les régressions linéaires sont les plus simples. Mais on exploite le plus souvent des méthodes de régression non linéaires. Une régression linéaire relève-t-elle de l'IA ? Sujet de débat. Au mieux en est-ce une forme des plus simples.

Il existe de nombreuses méthodes de régression, comme le **LASSO** (Least Absolute Shrinkage and Selection Operator)[98] qui date de 1986 et qui optimise notamment la sélection de variables exploitées pour calculer la régression non linéaire.

Les jeux de données en entrée comprennent plusieurs variables (x, y, z…). Il existe différentes formes de régression, notamment linéaire et non linéaire.

S'y ajoute aussi la notion d'overfitting et d'underfitting, qui décrit les méthodes de régression qui suivent plus ou moins de près les variations observées. Il faut éviter les deux et trouver le juste milieu ! C'est le travail des data scientists.

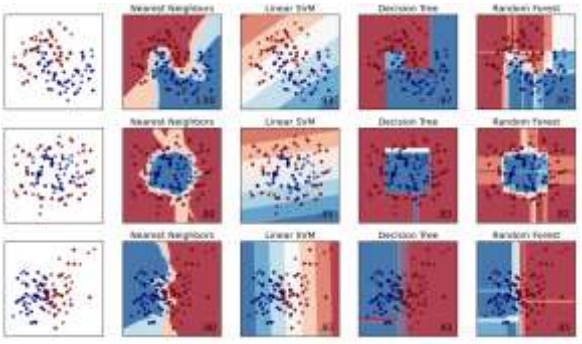


Les régressions peuvent être aussi réalisées avec des arbres de décision (CART), des modèles SVM, des réseaux de neurones, etc.

Clustering

Le clustering, le partitionnement ou la segmentation automatique est une méthode d'apprentissage non supervisé qui permet à partir d'un jeu de données non labellisé d'identifier des groupes de données proches les unes des autres, les clusters de données. Les méthodes de clustering permettent d'identifier les paramètres discriminants de ces différents segments ainsi que d'identifier de nouvelles classes d'objets par rapport à un jeu de classes existant.

Elles servent ensuite à prévoir l'appartenance à un segment d'une nouvelle donnée entrée dans le système. Là encore, si le clustering peut être automatisé, en mode non supervisé, le choix du modèle de clustering ne l'est pas nécessairement pour autant sauf avec des outils avancés comme ceux de DataRobot et Prevision.io.



source : http://mindcraft.ai/wp-content/uploads/2018/04/classifiers-comparison-scikit-learn-mod.png

La technique la plus répandue est l'algorithme des k-moyennes (*k-means*) qui vise à répartir les objets dans k classes d'objets distinctes. L'optimum de la segmentation est obtenu lorsque la distance moyenne entre les points de chaque classe et leur barycentre atteint un minimum.

Le machine learning à base de réseaux de neurones permet de son côté de segmenter des données avec une répartition quasi-arbitraire alors que les méthodes élémentaires ci-dessus sont limitées de ce point de vue-là.

[98] Voir 40 Techniques Used by Data Scientists, parVincent Granville, 2016.

Réduction de la dimensionnalité

La dimension des données devient de plus en plus grande à cause de la variété des big data. Un bon nombre d'algorithmes souffrent de la malédiction des grandes dimensions (« *curse of dimensionality* »).

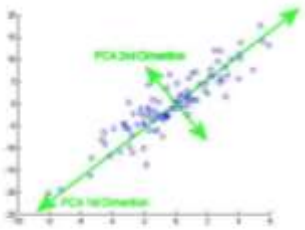
Il existe donc des techniques de nombreuses méthodes de réduction de dimension. Les plus classiques consistent à plonger les données (on parle d'*embedding*) dans un espace de plus faible dimension, de façon à préserver certaines propriétés.

Par exemple, l'Analyse en Composantes Principales (ou PCA, « *Principal Component Analysis* ») est une projection linéaire sur un espace, dont la dimension est le nombre souhaité de dimensions final, qui préserve le mieux la variance ou la dispersion des données.

Dans le cas de l'Analyse Discriminante (Linear Discriminant Analysis : LDA), on projette les données linéairement sur un espace, mais en essayant de préserver au mieux la discrimination entre les classes.

On pourra ainsi, par exemple, identifier les paramètres d'une segmentation client ou leur combinaison qui sont les plus pertinents pour prédire un comportement donné (churn, achat, …) ou pour identifier les paramètres clés qui permettent la détection d'un virus informatique. Cela permet de simplifier les modèles et améliore les prévisions dans la suite des opérations.

la complexité du machine learning augmente avec le nombre de dimensions ou variables examinées
paramètres de segmentation marketing clients, phénotype patients, état de capteurs d'une machine pour maintenance prédictive, …



on réduit le nombre de dimension avec la méthode PCA (Principal Component Analysis)
elle permet d'identifier les variables clés discriminantes

La réduction du nombre de variables utilisées va aussi réduire la consommation de ressources machines. Mais attention, les variables discriminantes ou facteurs de corrélation ne sont pas forcément des facteurs de causalité. Ces derniers peuvent être externes aux variables analysées[99] !

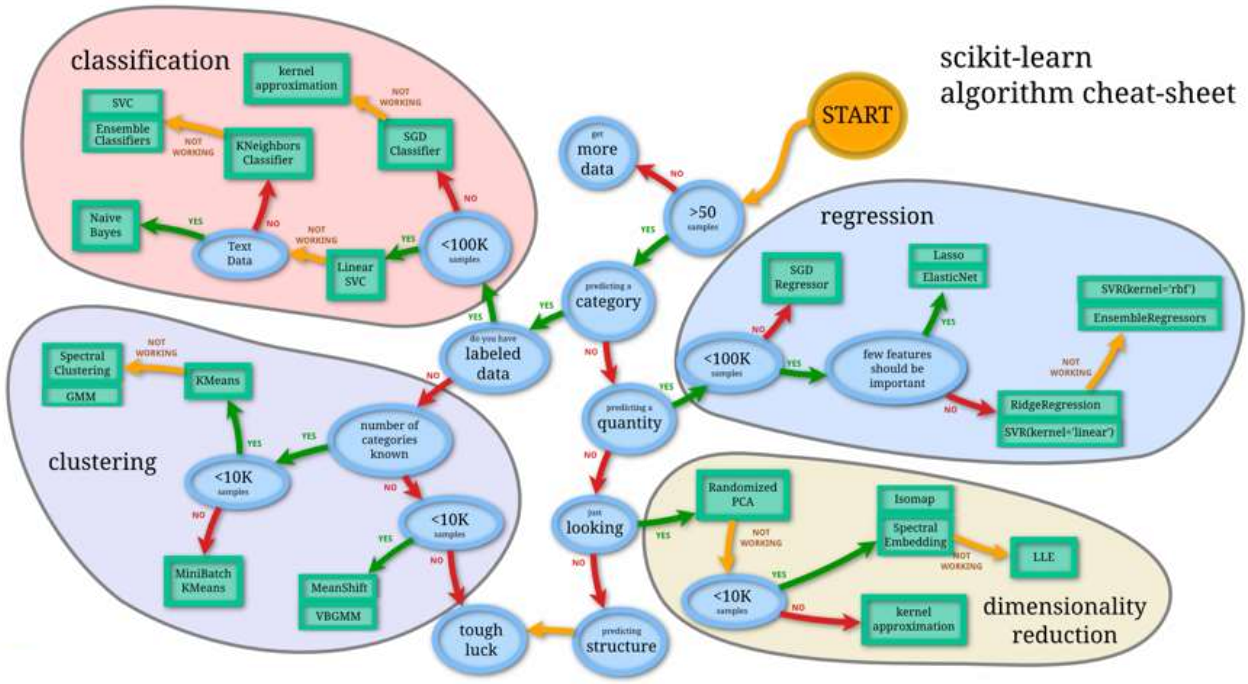
Les techniques de réduction de dimension, et notamment la PCA, sont très largement utilisées dans le machine learning et le deep learning (qui a sa propre version de la PCA calculée par un réseau de neurones, dit auto-encoder).

Outils du machine learning

Le machine learning nécessite d'abord de bien déterminer la typologie du problème à résoudre et des données disponibles. Le schéma *ci-dessous* associé à la bibliothèque d'origine française **scikit-learn** développée par Inria est un exemple d'arbre de décision permettant de déterminer la méthode à utiliser en fonction du problème à résoudre : régression, classification, clustering ou réduction des dimensions et les sous-méthodesen fonction de la typologie et du volume des données. C'est le b-a-ba du data scientist.

La panoplie des outils de machine learning est de fait abondante selon leur niveau de personnalisation et d'accès à des publics plus ou moins techniques.

[99] C'est très bien expliqué dans cette tribune de Laurent Alexandre parue en novembre 2018 dans L'Express : Posséder un Picasso protégerait du cancer.



On peut les classifier en plusieurs catégories :

- Des **langages de programmation** comme Python, Java, C++ ou autres qui sont utilisés conjointement avec des bibliothèques de calcul spécialisées dans le machine learning. Il y a aussi le langage Julia associé aux bibliothèques JuliaStats qui permettent de créer des applications statistiques et de machine learning[100].

- Des **bibliothèques associées** qui permettent de développer les modèles d'apprentissage ou d'auto-apprentissage et de les mettre ensuite en production. Ces outils tournent sur poste de travail et dans le cloud[101]. On y trouve notamment scikit-learn, qui est d'origine française

environnements de travail	bibliothèques
Apache Zeppelin	Scikit-Learn / Python
PyCharm	TensorFlow
Azure Machine Learning Studio	Mlpack / C++
Amazon Machine Learning	RapidMiner / Java
Google Cloud Machine Learning	Weka / Java
	Spark MLLib / Scala
	Torch / Lua
	JuliaStats / Julia

- Des **environnements de travail,** ou IDE pour Integrated Development Environment, qui permettent de paramétrer ses systèmes et de visualiser les résultats, souvent de manière graphique. Ils servent à tester différentes méthodes de classification, régression et clustering pour définir les modèles à appliquer. Ils peuvent aussi servir à piloter la mise en production des solutions retenues.

[100] Le langage Julia date de 2012. Voir Julia pourrait accélérer le développement d'app nécessitant des mathématiques complexes, Services Mobiles, août 2019.

[101] On peut y ajouter **Nvidia Rapids** qui est un ensemble de bibliothèques open source en Python dédiées au machine learning et s'appuyant sur les APIs CUDA-X qui supportent la distribution de traitements sur plusieurs GPU et serveurs. Elle comprend notamment cuDF (manipulation de données), cuML (bibliothèques de machine learning scikit-learn exécutables sur GPU) et cuGraph (pour la gestion de graphes).

Parmi eux, les solutions d'**IBM**, de **SAS**, de **Knime** (2008, Suisse, 20M€), de **RapidMiner** (2007, USA, $36M)**,** les solutions de **Cognitive Scale** (2013, USA, $40M[102]), le Data Science Workbench de **Cloudera** (2008, USA, $1B) et le Data Studio de **Dataiku** (2013, France, $246,8M). Il y a enfin **Matlab**. Cette catégorie comprend de nombreux autres acteurs tels que **Alteryx** (2010, USA, $163M), **Predixion Software** (2009, USA, $37M), **Alpine Data Labs** (2011, USA, $25M), **H2O.ai** (2012, USA, $146M) et **Lavastorm** (1999, USA, $55M).

- Des **outils d'automatisation** de la recherche de méthodes d'apprentissage aussi appelés AutoML[103] comme chez **DataRobot** (2012, $225M, vidéo), **Prevision.io** (2016, France, 7,5M€, vidéo et démonstration), **Synaplus** (2014, France) et son outil Cozmo, **DataValoris** (2015, France) et RAISE, le Coréen **Daria** (2015, $1,1M), **Solidware** (2014, Corée du Sud et créé par des français) propose Davinci Labs qui gère de manière intégrée tout le cycle de la création de solutions de machine learning. Enfin, Darwin de **SparkCognition** (2014, USA, $56,3M) est un outil équivalent. **Peltarion** (2005, Suède, $36,8M) propose aussi un outil en cloud qui gère tout le cycle, de l'ingestion des données jusqu'à la production. **MyDataModels** (2018, France, 3,5M€) propose aussi un environnement d'analyse des données de type AutoML destiné aux chercheurs et ingénieurs. **Google AutoML** peut aussi servir à analyser ses données, en plus de ses capacités dans le traitement de l'image et du langage. Ces outils récupèrent les données du client. Ils permettent de tester diverses méthodes d'apprentissage relevant du machine learning pour trouver celles qui sont les plus pertinentes par rapport à un objectif à atteindre, de manière plus ou moins automatique selon les cas. Ils exécutent les tests de modèles en parallèle – parfois sur différentes ressources dans le cloud - pour prédire les valeurs d'une variable dans un tableau à partir d'un tableau d'entraînement[104].

[102] Le marketing produit de Cognitive Scale est caricatural : il n'est franchement pas évidement de comprendre ce que réalise le produit. Celui de DataRobot est bien mieux réalisé et clair.

[103] Voici une description des différentes méthodes et niveaux d'automatisation de la creation de modèles de machine learning : What is neural architecture search? An overview of NAS and a discussion on how it compares to hyperparameter optimization par Liam Li et Ameet Talwalkar, décembre 2018. Voir aussi 6 Top AutoML Frameworks for Machine Learning Applications (May 2019) par Alibaba, septembre 2019.

[104] Les outils permettent de se passer de programmation. Prevision.io créé un modèle prêt à l'emploi sans programmation et qui sera aussi exploitable par du code dans une application spécifique via une API en cloud. Voir ce descriptif précis du mode opératoire de prevision.io : Building a production-ready machine learning system with Prevision.io de Gerome Pistre, octobre 2017.

Dans la même lignée, Amazon **AutoGluon** simplifie le choix des modèles et des hyperparamètres pour les développeurs de solutions de machine learning ainsi que de deep learning[105].

- Des **outils destinés aux utilisateurs** pour leur permettre d'analyser leurs données et de produire des rapports graphiques pertinents en se passant théoriquement de data scientists. C'est ce que propose **Thoughtspot** (2012, USA, $150,7M) avec une solution qui utilise le machine learning pour identifier les besoins de présentation des données de l'utilisateur[106]. **Invenis** (2015, France), **Mondobrain** (2014, USA, crée par le Français Augustin Huret, $13,3M, vidéo) et **Jet-Pack Data** (2016, France) sont sur le même créneau, ce dernier générant automatiquement un dashboard multicritères à partir d'un jeu de données arbitraire[107]. **Tableau** (2003, USA, $15M, IPO en 2013, acquis par Salesforce en 2019) propose une sorte de Business Objects en cloud multiécrans et doté de fonctions collaboratives. Ils ont fait l'acquisition de la startup **Empirical Systems** (2015, USA, $2,5M) en juin 2018 pour ajouter à leur plateforme l'automatisation de la création de modèles de présentation de données. Bref, de faire du machine learning plus ou moins automatique. Il faut y ajouter **Dataiku** (2013, France, $246,8M)[108]. On peut aussi y ranger **Palantir** (2004, USA, $20B).
- Des outils pour la création de solutions de machine learning pour les objets connectés, comme ceux de **Numericcal** (2016, USA). Ils optimisent le code généré pour tourner dans des systèmes contraints par la consommation d'énergie, la mémoire et la puissance disponibles.

Les compétences nécessaires pour créer des solutions de machine learning sont multiples. En amont, elles relèvent de la collecte et de l'organisation des données. C'est le big data. En son cœur, elle relève de la data science et des data scientists, qui exploitent ces données avec les logiciels du machine learning. Enfin, en aval les développeurs traditionnels continuent de créer des solutions logicielles exploitables par les utilisateurs des entreprises ou le grand public.

Une bonne solution de machine learning doit être alimentée par des sources de données adaptées au problème à résoudre. Ces données doivent contenir suffisamment d'informations à valeur statistiques permettant de faire des régressions, segmentations ou prévisions. Leur bonne distribution spatiale dans l'univers du possible qui est étudié est encore plus importante que leur précision à l'échelle unitaire.

Réseaux de neurones

Les réseaux de neurones visent à reproduire approximativement par bio-mimétisme le fonctionnement des neurones biologiques avec des sous-ensembles matériels et logiciels capables de faire des calculs à partir de données en entrée et de générer un résultat en sortie. C'est une technique utilisée dans le machine learning et dans sa variante avancée du deep learning.

Les neurones artificiels

Un neurone artificiel est un objet logiciel qui récupère des variables numériques en entrée (x_1, …, x_n) associées à un poids (w_1, …, w_n) et combine ces valeurs pour générer une valeur en sortie.

---

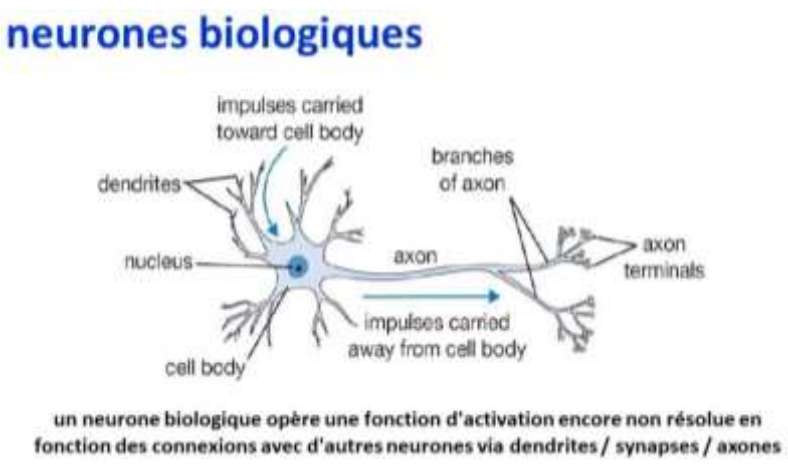
[105] Voir Avec AutoGluon, Amazon automatise le deep learning par Paul Krill, janvier 2020.

[106] Le montant levé par ThoughtSpot n'est pas très étonnant car la startup vise un marché de volume, les utilisateurs et apporte beaucoup de valeur à ses clients d'entreprises.

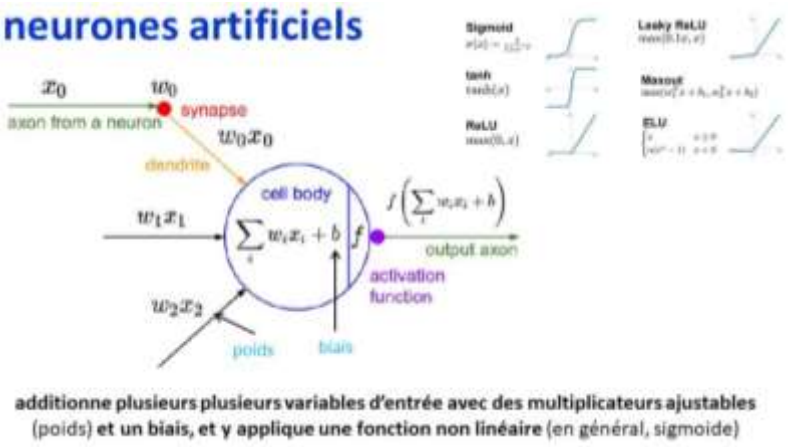
[107] Il s'agit presque toujours d'un tableau avec plusieurs colonnes de paramètres, la dernière colonne étant un « tag » déterminant d'un résultat, par exemple, un taux de défaut de fabrications de pièces, du churn client, ou tout autre comportement de machine ou d'individu.

[108] Dataiku lançait en 2018 une nouvelle version de son Data Science Studio s'adressant aux utilisateurs qui peuvent ainsi prototyper, créer, déployer et gérer leurs modèles de données grâce à une interface graphique accessible et personnalisable. Des astuces telles que la fonction « Live Model Competition » permettent d'y comparer des modèles en temps réel sans attendre la fin des calculs de création du modèle

Le neurone artificiel moderne fait la somme des entrées multipliées par leur poids, additionne un biais (b qui qui permet de s'assurer que le résultat reste entre 0 et 1) et lui applique ensuite une **fonction d'activation** qui est une fonction non linéaire comme une sigmoïde qui génère une valeur comprise entre 0 et 1, ou entre -1 et +1 générant une valeur % statistique facile à exploiter dans le reste du réseau de neurones[109].



Les sigmoïdes sont utilisées dans les réseaux fully connected et les ReLU (qui met à zéro les valeurs négatives) le sont dans les réseaux convolutifs que nous verrons plus loin. La non linéarité de la fonction d'activation est une caractéristique clé des réseaux de neurones depuis les années 1980[110]. La non-linéarité permet de faire converger le réseau de neurones lors de son entraînement par rétropropagation des gradients.



Elle sert à filtrer le signal, pour ne conserver que les informations utiles et supprimer le bruit inutile.

Le procédé imite vaguement le fonctionnement d'un neurone biologique qui est largement plus complexe et dont le fonctionnement dépend d'un très grand nombre de paramètres biochimiques.

Un neurone isolé ne sert pas à grand-chose[111]. Ils sont assemblés dans des réseaux de neurones. Un réseau de neurones de machine learning comprend souvent plusieurs couches de neurones. Les neurones d'une même couche ne sont généralement pas connectés entre eux contrairement aux neurones du cortex, histoire de créer des systèmes plus simples. Ils sont connectés aux neurones de la couche suivante.

---

[109] Voir cet excellent article qui décrit les variantes de fonctions d'activation avec notamment les softmax : Activation Functions : Sigmoid, ReLU, Leaky ReLU and Softmax basics for Neural Networks and Deep Learning par Himanshu Sharma, janvier 2019.

[110] Le passage des neurones basiques de type perceptron aux neurones exploitant des fonctions non linéaires date du début des années 1980. Yann Le Cun raconte que cela a été permis par l'avènement de stations de travail Sun dotées de processeurs capables de calculs en nombres flottants. On a aussi oublié que le processeur du premier IBM PC, l'Intel 8088 était complété d'un coprocesseur de calculs en nombre flottants, le 8087, lancé en 1980. Il a été suivi du 80287 qui accompagnait le processeur 16 bits 80286 puis du 80387 qui complétait le processeur 32 bits 80386 lancé en 1985. A partir des Pentium lancés en 1993, Intel a intégré les calculs flottants dans le CPU. Le passage au calcul flottant qui a permis la création de réseaux de neurones multicouches et leur entraînement.

[111] Un neurone du cortex cérébral est généralement relié par son axone à des milliers d'autres neurones via plusieurs synapses qui s'associent à une dendrite, une sorte d'excroissance de neurone. Il y a huit neurotransmetteurs différents qui font fonctionner les synapses. Et l'ensemble est régulé par l'expression de 6000 gènes différents dans les neurones et par des cellules gliales qui alimentent les neurones en énergie et qui régulent la production de neurotransmetteurs et la conductivité des axones via la myéline qui les entoure. Bref, c'est très compliqué ! Mais on en découvre tous les jours sur la complexité des neurones ! Voir Surprise! Neurons are Now More Complex than We Thought, de Carlos Perez, 2018. Et les milliers de microtubules qui constituent la structure des neurones pourraient elles-mêmes jouer un rôle clé dans la mémoire.

On évite généralement des connexions circulaires entre neurones pour éviter de faire fonctionner le réseau en boucle lors de son apprentissage, sauf dans le cas des réseaux récurrents. C'est en tout cas vrai pour les réseaux de neurones convolutifs que nous verrons plus loin.

Une couche cachée permet de générer une méthode de classification non linéaire complexe. On parle de deep learning lorsque le réseau de neurones comprend plus d'une couche cachée. C'est pour cela que le deep learning est considéré comme étant un sous-ensemble du machine learning[112].

La « connaissance » du réseau de neurones est acquise via un processus d'apprentissage permettant d'ajuster le poids des interconnexions entre neurones pour que les objets en entrée du réseau de neurones soient reconnus en sortie, en général avec un label descriptif, aussi appelé une classe, ou une valeur, comme la dénomination d'un objet ou le nom d'une personne pour une image en entrée. Il s'agit d'une connaissance purement probabiliste. La réponse n'est pas déterministe. Elle est généralement exprimée sous la forme d'un taux de probabilité, normalement supérieur à 90%.

Le savoir intégré dans un réseau de neurones n'est pas symbolique. Il ne sait pas donner de sens aux objets qu'il détecte ou aux calculs qu'il réalise ni expliquer les raisons de sa décision.

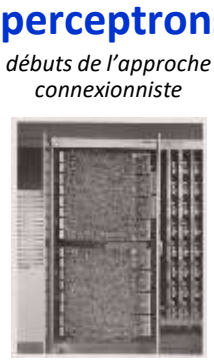
Dans l'informatique classique, on classifie les processeurs en deux catégories : les CISC à jeu d'instruction complexe (comme chez Intel) et les RISC à jeu d'instruction simple (comme chez arm ou RISC-V). Un processeur mettant en oeuvre un réseau de neurones est classifié en NISC : « no instruction set computer » car le système ne fonctionne pas avec une unité de contrôle enchaînant des instructions à la suite comme dans un modèle de Von Neumann.

Les perceptrons

Le concept des réseaux de neurones a vu le jour en 1943 avec les travaux de **Warren McCullochs** et **Walter Pitts**. En 1949, **Donald Hebb** ajouta le principe de modulation des connexions entre neurones, permettant aux neurones de mémoriser de l'expérience.

Le premier réseau de neurones matériel fut créé par **Marvin Minsky** et **Dean Edmons** en 1950 alors qu'ils étaient étudiants à Harvard. Le SNARC simulait 40 neurones basiques avec 3000 lampes à tubes de type triodes ! Et c'était avant le Summer Camp de Darmouth de 1956 !

Frank Rosenblatt, un collègue de Marvin Minsky, créa ensuite le concept du **perceptron** en 1957 qui était un neurone assez simple dans son principe avec une fonction de transfert binaire, générant un 0 ou un 1 en sortie. Le premier perceptron était donc un réseau de neurones artificiels à une seule couche tournant sous forme de logiciel dans un **IBM 704**, le premier ordinateur du constructeur doté de mémoires à tores magnétiques.

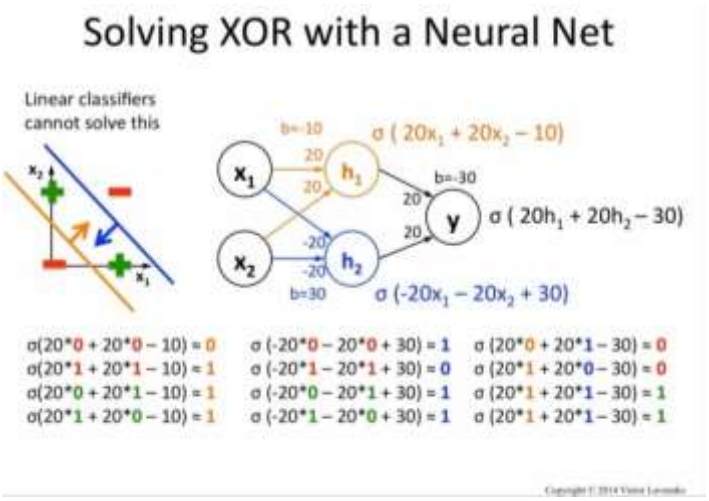


C'était un outil de classification linéaire utilisant un seul extracteur de caractéristique et sortant une simple valeur binaire. Un calculateur dédié fut ensuite construit en 1957, le "Mark 1 perceptron", pour faire de la reconnaissance d'images avec 400 cellules photo.

[112] Le deep learning est dénommé apprentissage profond en français mais j'utilise l'appellation anglaise dans ce document.

Le poids de neurones était géré par des potentiomètres motorisés. En 1960, une version plus sophistiquée était créée, **ADALINE** (Adaptive Linear Neuron or later Adaptive Linear Element) un réseau de neurones monocouche construit avec des memristors par **Bernard Widrow** à Stanford.

En 1969, **Marvin Minsky** publia avec **Seymour Papert** le livre **Perceptrons** qui critiquait les travaux de Frank Rosenblatt et sur un point très spécifique portant sur l'impossibilité de coder une porte logique XOR avec un perceptron.



Une porte XOR détecte si les deux entrées binaires sont identiques : 0, 0 et 1, 1 deviennent 1 et 0,1 ou 1, 0 deviennent 0. Tout en proposant une solution de contournement associant deux couches de neurones pour mettre en œuvre la porte XOR. Le livre n'était donc pas si destructif que cela !

C'était même la voie vers les réseaux multi-couches qui, en effet, peuvent calculer un XOR. Mais il faudra attendre près de 20 ans pour qu'ils voient le jour.

Les auteurs contribuèrent cependant à mettre un coup d'arrêt à ces développements, le coup de grâce arrivant avec le **rapport Lighthill** publié au Royaume Uni en 1973. Cela fit perdre un temps considérable à l'ensemble des recherches en IA, ce d'autant plus que les réseaux neuronaux sont devenus, depuis, un pan fondamental des progrès dans tous les étages de l'IA. Marvin Minsky reconnut toutefois son erreur d'appréciation dans les années 1980, après le décès de Frank Rosenblatt.

Du machine learning au deep learning

Les réseaux neuronaux ont connu ensuite un fort développement à partir de 2012 et dans leur mise en œuvre d'abord dans le machine learning puis avec le deep learning, qui exploite des réseaux de neurones avec un grand nombre de couches. C'est pour cela qu'ils sont profonds !

Dans le machine learning, les réseaux de neurones à une ou deux couches cachées permettent de créer des méthodes de classification d'objets plus sophistiquées et de données comportant un grand nombre de dimensions comme les pixels d'une image.

De nombreuses méthodes d'organisation de réseaux de neurones sophistiqués sont ensuite apparues pour permettre la reconnaissance de la parole et d'images. Elles sont évoquées dans la partie sur le deep learning.

Enfin, citons les réseaux de neurones multimodes qui exploitent des sources d'informations complémentaires, classiquement, de l'audio et de la vidéo, pour améliorer la qualité de la captation. L'audio d'une vidéo permet par exemple d'améliorer la capacité à tagger le contenu de la vidéo. Cela peut aller jusqu'à lire sur les lèvres pour améliorer la reconnaissance de la parole.

L'imagerie 2D complétée par des informations de profondeur améliorera la capacité de détection d'objets complexes. La vidéo d'un visage permettra d'améliorer la captation de la parole par l'équivalent numérique de la lecture sur les lèvres.

Un exemple classique de réseau de neurones simple est celui de l'évaluation du prix de vente potentiel d'un appartement en fonction de quelques critères clés discriminants comme sa surface, son âge et son ancienneté.

Avec quelques paramètres numériques de ce type, un tel réseau peut se contenter de n'avoir que quelques couches, deux dans l'exemple. En sortie de réseau, il générera une estimation du prix de l'appartement[113].

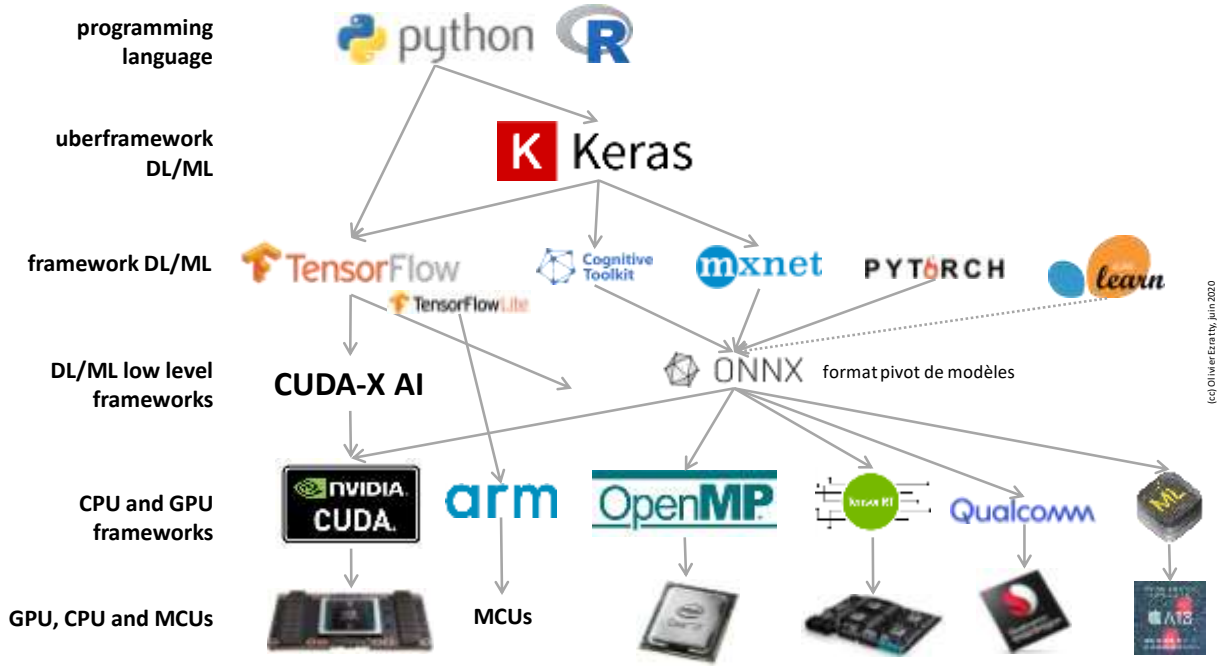
Ce sont des réseaux multicouches dits **feed forward** : on les alimente en amont avec des données qui rentrent dans les neurones de la première couche puis passent aux neurones de la couche suivante via leurs synapses, ainsi de suite jusqu'à la dernière couche qui donne une réponse.

Sur les schémas, l'information circule de gauche à droite pendant l'exécution du réseau de neurones. On appelle aussi cela une inférence.

Comment entraîne-t-on un réseau de neurones, à savoir, comment ajuste-t-on le poids de chacune des synapses de chaque neurone du réseau ? Nous répondrons à cette question dans la partie consacrée au deep learning et à la rétropropagation d'erreurs[114].

Programmation de réseaux de neurones

D'un point de vue pratique, la programmation de réseaux de neurones s'appuie sur des bibliothèques logicielles spécialisées comme **cuDNN**, **MKL** ou **OpenNN**. On peu aussi citer **Synaptic** qui est une bibliothèque utilisable avec node.js dans un navigateur en JavaScript.



[113] Une solution voisine de cet exemple académique semble être maintenant opérationnelle sur le site SeLoger. Voir Dans les coulisses du nouvel outil d'estimation de prix de SeLoger de Justine Gay, octobre 2018.

[114] Voir Neural Networks and Deep Learning par Michael Nielsen, 2018 (224 pages) qui explique bien les basiques mathématiques des réseaux de neurones, de leur entraînement, y compris dans le cas du deep learning.

Ces bibliothèques de réseaux de neurones sont souvent exploitées elles-mêmes par des bibliothèques de machine learning ou de deep learning, comme **TensorFlow**, qui masquent la complexité du pilotage de réseaux de neurones à bas niveau et permettent par exemple de définir les modèles de réseaux de neurones convolutifs de reconnaissance d'images et de les entraîner.

C'est illustré dans le schéma *ci-dessus* qui empile les couches utilisées dans le développement de solutions d'IA avec un développement comprenant un framework d'abstraction élevé utilisant un framework, comme le framework **Keras** qui se situe au-dessus de TensorFlow, puis une bibliothèque de réseau de neurones, suivie d'une bibliothèque de pilotage de GPU comme CUDA chez Nvidia, et enfin, un GPU ou un CPU au niveau matériel.

Les frameworks **CNTK** de Microsoft**, PyTorch** de Facebook supportent pour leur part un format de description de modèle intermédiaire ONNX qui supporte de son côté les principaux frameworks de CPU et GPU du marché.

TensorRT est le framework d'exécution de modèles entraînés de Nvidia. **Qualcomm** a le sien pour ses Snapdragon, le Qualcomm Neural Processing SDK for AI. **CoreML** est de son côté lié aux plateformes Apple[115].

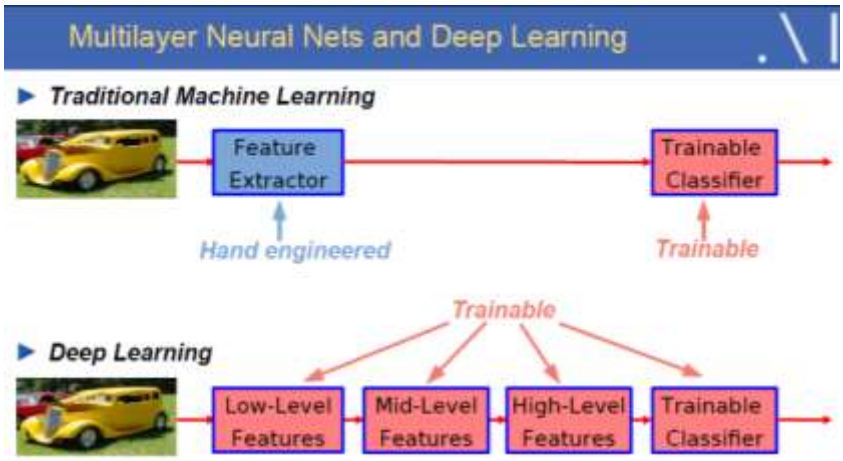
Deep learning

Le deep learning est un sous-ensemble des techniques de machine learning à base de réseaux de neurones qui s'appuient sur des réseaux de neurones à plusieurs couches dites cachées.

Celles-ci permettent par exemple de décomposer de manière hiérarchique le contenu d'une donnée complexe comme de la voix ou une image pour la classifier ensuite : identifier des mots pour la voix ou associer des tags descriptifs à des images.

C'est le principe de l'une des grandes catégories de réseaux de neurones de deep learning, les réseaux convolutifs (schéma *ci-dessous*). Un réseau peut être profond mais aussi large si le nombre de neurones est élevé dans chaque couche.

Le deep learning remplace les méthodes antérieures du machine learning à base de « handcraft features » qui consistaient à définir à la main les éléments à rechercher dans les objets (formes dans les images, tournures dans les textes)[116]. Le schéma *ci-contre*[117] illustre bien cette différence entre machine learning avec extraction manuelle de features et deep learning qui gère automatiquement différents niveaux d'abstraction de ces features.

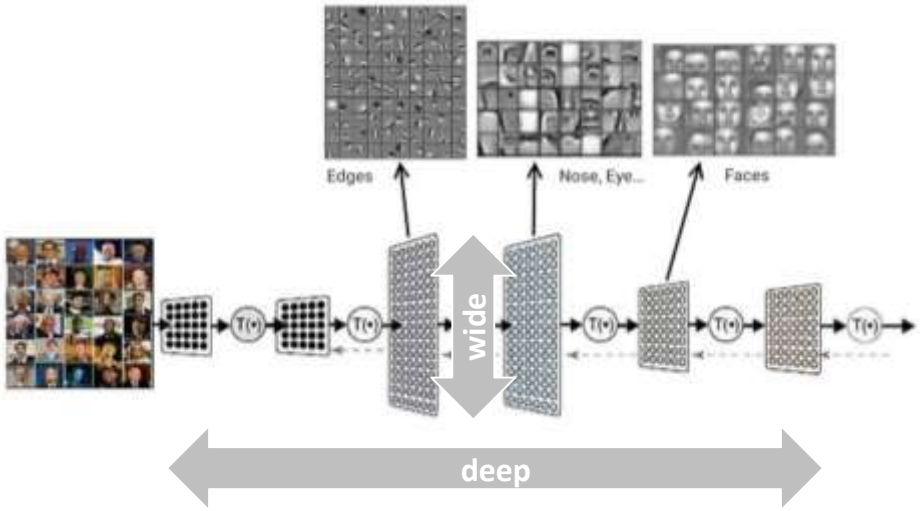


[115] Voir La version 6 du langage Swift ciblera le machine learning par Paul Krill, février 2020 qui évoque le fait que le langage de développement d'applications mobiles d'Apple supporte maintenant le développement de solutions exploitant du machine learning et s'appuyant notamment sur CoreML.

[116] Voir Convolutional Neural Networks par Christof Angermueller et Alex Kendall (48 slides).

[117] Le schéma provient de l'excellente présentation The Power and Limits Of Deep Learning de Yann Le Cun à Harvard, mars 2019 (120 slides).

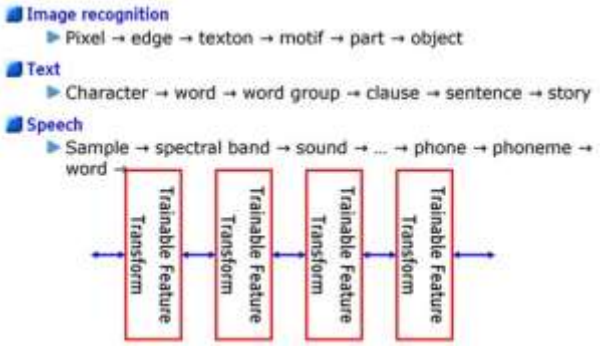
Dans le deep learning, notamment pour la détection d'images, le réseau de neurones découvre tout seul ces composantes avec des niveaux d'abstraction évoluant de bas en haut et de couche en couche[118].

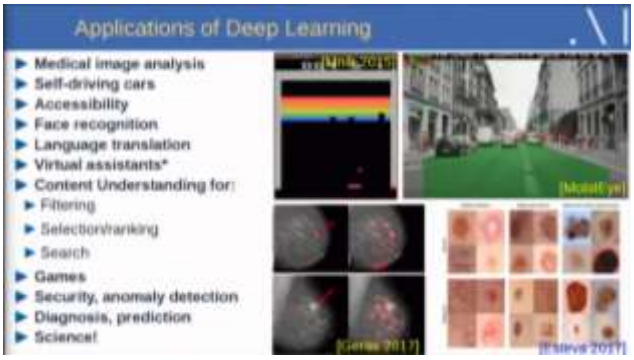


chaque niveau d'un réseau de neurones profond réalise une fonction de classification d'un type d'objet en un type d'objet supérieur
(pixels => forme => nez => visage => individu)

un réseau de neurones profond de type convolutionnel comprend plusieurs couches "cachées" qui transforment les données en entrée en données ayant un niveau d'abstraction supérieur

Le deep learning sert le plus souvent au traitement du langage, de la parole, du bruit, de l'écriture, des images et, c'est moins connu, de graphes. Il a d'autres usages dans les outils d'aide à la décision, dans les jeux tels que le Go avec AlphaGo et même dans l'exploitation de données structurées, dans la cybersécurité et d'une manière générale dans la recherche scientifique[119] comme en génomique.





Le deep learning permet aussi de générer des contenus artificiels, extrapolés à partir de contenus réels, notamment des images, que nous verrons aussi, et qui s'appuient sur des modèles génératifs et « adversariaux » (ou GAN, pour Generative Adversarial Networks).

Il ne faut pas confondre la notion de deep learning avec les différents modèles d'entraînement qui s'appliquent aussi bien au deep learning qu'au machine learning : l'entraînement supervisé, par renforcement ou auto-supervisé. Certaines extensions du deep learning comme la programmation différentielle et certains modèles de graphes lui permettent d'adapter dynamiquement la structure du réseau de neurones.

---

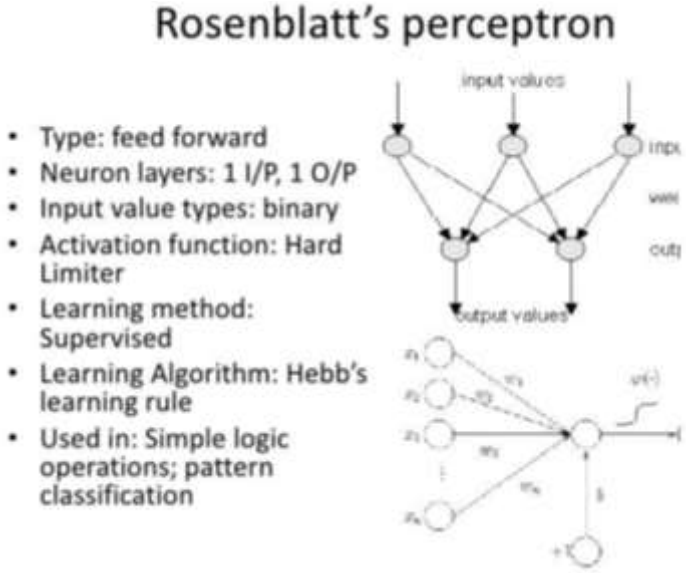
[118] Pour comprendre le fonctionnement du deep learning, vous pouvez profiter de cette excellente présentation : Deep Learning : A crash Course par Andrew Glassner, 2018 (3h33mn). Andrew Glassner est un spécialiste de l'informatique graphique et l'auteur de Deep Learning From Basics to Practice en deux volumes (2018).

[119] Le slide de droite est issu de la conférence de Yann Le Cun à l'USI à Paris en juin 2018 (vidéo). Celui de gauche provient également de Yann Le Cun, mais de sa conférence inaugurale au Collège de France en 2016 (lien).

Par contre, le deep learning n'est pas la panacée. Il s'appuie sur des modèles probabilistes comme le machine learning dont il est une classe de variantes. Il n'est pour l'instant pas adapté au raisonnement, à la génération de sens commun, à la création de robots intelligents et agiles et d'une manière générale à la création d'IA générales (AGI, dont nous parlerons plus tard). Cela pourrait cependant changer un jour.

Évolutions du deep learning

Les outils de deep learning s'appuient sur différentes variantes de réseaux de neurones pour leur mise en œuvre pratique. Leur histoire remonte aux perceptrons de **Franck Rosenblatt** de 1957. L'histoire du deep learning a véritablement démarré plus de 20 ans plus tard, dans les années 1980.



perceptrons 1957
multi-layered perceptron 1969
back propagation 1974
recurrent neural networks 1982
neocognitrons 1983
error propagation 1986
restricted boltzmann machine 1986
time delay neural networks 1989
forward propagation 199X
convolutional neural networks 1998
deep belief networks 2006
stacked autoencoders 2007
AlexNet / google imagenet 2012
captule networks 2017
federated learning 2017
transformers 2017

Il a cependant fallu attendre 1995 pour que l'on puisse les mettre en œuvre en pratique, sans doute, grâce aux progrès matériels, à la loi de Moore mais aussi aux progrès conceptuels, notamment aux travaux d'Alexander Weibel en 1989, Yann Le Cun en 1988 et 1998 et à Geoff Hinton, particulièrement à partir de 1986. Le premier est le père des réseaux TDNN de reconnaissance de phonèmes, le second des réseaux convolutifs à rétropropagation d'erreurs tandis que le dernier est considéré comme étant le père de nombreux concepts de base du deep learning. Geoff Hinton, Yann Le Cun et Yoshua Bengio ont reçu ensemble la médaille de Turing en mars 2019 pour leurs travaux sur le deep learning[120].

Il est de bon ton de déclarer que les chercheurs n'ont pas produit grand-chose depuis et que le deep learning doit tout aux progrès du matériel et à l'abondance de données pour entraîner les systèmes. Quand on y regarde de plus près, on se rend compte qu'au contraire, les chercheurs n'ont pas cessé de faire avancer le domaine[121]. Et d'année en année, des progrès conceptuels et pratiques font avancer les réseaux de neurones et le deep learning, ne serait-ce qu'avec les réseaux génératifs mais aussi avec les transformers et la notion de compositionnalité que nous verrons rapidement plus loin.

Le champ du deep learning est en perpétuelle évolution. On voit fleurir régulièrement de nouveaux types de réseaux de neurones, que ce soit pour la reconnaissance d'images (identification d'objets, segmentation graphique d'objets, labellisation d'objets, temps réel) ou dans le traitement du langage (traduction, questions/réponses, représentation des connaissances) sans compter le fascinant champ des réseaux de neurones génératifs qui créent des contenus nouveaux à partir de contenus existants.

---

[120] Geoff Hinton raconte les débuts du deep learning dans la conférence 'The Deep Learning Revolution' - Geoffrey Hinton - RSE President's Lecture 2019, juillet 2019 (1h23). Et notamment le rejet de l'establishment de l'IA jusqu'en 2009.

[121] Voir par exemple Beyond backpropagation : can we go deeper than deep learning ? de Marina Yao, novembre 2017.

Chaque année, un nouveau réseau de neurones rend obsolète ceux de l'année précédente. C'est un monde de remise en cause permanente de l'état de l'art.

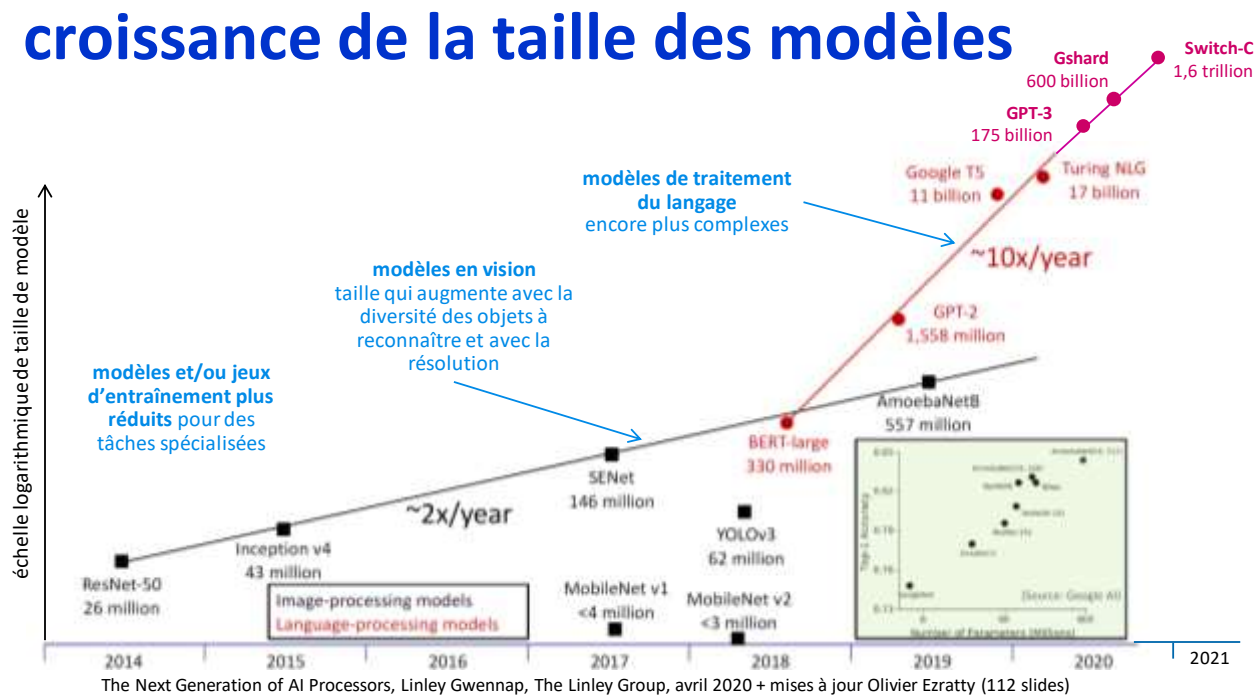
C'est ce que nous allons voir dans ce qui suit. Ces avancées du deep learning sont étalées sur plusieurs décennies et sont continues. Elles sont évoquées ici de manière chronologique selon leur date d'apparition. Si vous souhaitez aller directement à l'essentiel, consultez en priorité les réseaux convolutifs, les réseaux de neurones récurrents et à mémoire et enfin, les réseaux génératifs.

Rétropropagation d'erreurs (1969)

Un réseau de neurones a besoin d'être entraîné, à savoir que le poids des synapses a besoin d'être ajusté pour que le réseau de neurones génère de bons résultats. Or ces paramètres sont très nombreux dans un réseau de neurones de deep learning. Ils peuvent être plusieurs milliards ! Le record fin 2020 était le réseau de neurones de traduction **GShard** avec ses 600 milliards de paramètres[122]. Il atteignait 1,6 trillions de paramètres (au sens américain du terme) avec le réseau de neurones Switch-C de Google Brain[123].

Comment fait-on donc pour définir le poids de ces synapses ? La méthode la plus courante consiste à utiliser la rétropropagation du gradient.

Elle fonctionne couche par couche en partant du résultat et en ajustant le poids des neurones pour permettre au réseau d'identifier les objets de la base d'entraînement fournis en entrée. Cette rétropropagation fonctionne en ajustant un par un les poids des neurones de chaque couche et en scannant un par un les objets du jeu de test pour optimiser le taux de reconnaissance, en minimisant ce que l'on appelle la « fonction d'erreur », soit la différence entre ce que génère le réseau pendant sa phase d'entraînement et la bonne réponse dont on dispose déjà dans la base d'entraînement.



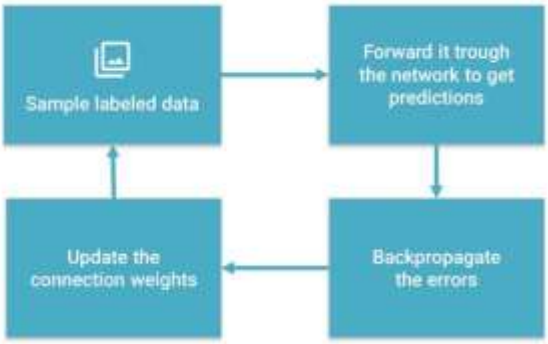
The Next Generation of AI Processors, Linley Gwennap, The Linley Group, avril 2020 + mises à jour Olivier Ezratty (112 slides)

[122] Voir GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding par Dmitry Lepikhin, juin 2020 (35 pages). C'est un réseau de neurones de 600 milliards de paramètres pour faire de la traduction entraîné sur 2048 TPU de Google pendant quatre jours. Il peut traduire des textes de 100 langages en anglais. Mais cela ne va pas s'arrêter là. La technique ZeRO permettrait d'optimiser l'entraînement de ce genre de réseau de neurones pour dépasser le trillion de paramètres. Voir ZeRO: Memory Optimizations Toward Training Trillion Parameter Models par Samyam Rajbhandari et al, mai 2020 (24 pages).

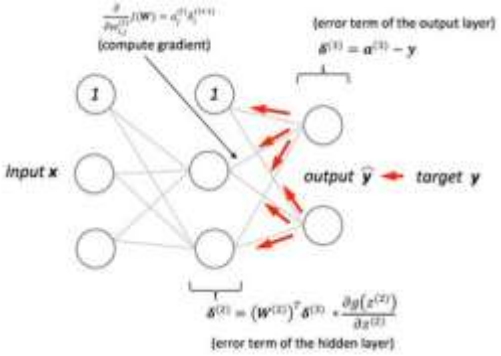
[123] Voir Switch transformers: scaling to trillion parameter models with simple and efficient sparsity par William Fedus et al, 2021 (31 pages).

L'apprentissage des réseaux de neurones est généralement supervisé et automatique ! Supervisé car il utilise des labels descriptifs des objets d'une base de référence et automatique car les poids synaptiques des neurones sont ajustés automatiquement grâce à ces méthodes de rétropropagation programmées dans le système d'entraînement.

Les évolutions des méthodes de rétropropagation créées par la suite visaient surtout à économiser du temps machine car l'opération est très fastidieuse puisqu'elle doit être répétée pour chaque neurone du réseau et pour chaque objet de la base de référence. Cela donne une combinatoire très élevée !



l'apprentissage supervisé avec rétro-propagation d'erreurs mesure la différence entre prédictions et valeurs souhaitées, celle-ci servant à modifier les poids des neurones pour améliorer les prédictions



pour chaque objet testé, l'erreur est répercutées dans les neurones amont en modifiant les poids des synapses au prorata de leur poids respectif

La méthode est perfectionnée en 1986 par David Rumelhart, Geoff Hinton et Ronald Williams[124] alors qu'au même moment, Yann Le Cun publiait une version de l'algorithme dans sa thèse, en 1987. La plus couramment utilisée aujourd'hui est la **descente stochastique de gradient** (ou SGD pour stochastic gradient descent en anglais), vue un peu plus loin, et qui permet d'améliorer la vitesse de convergence des réseaux lors de leur entraînement. Là encore, ce sont les capacités de calcul en nombres flottants qui ont permis à ces techniques de proliférer.

La *backprop* a ceci de particulier qu'elle ne relève pas a priori du biomimétisme. Ou tout du moins, on n'en sait rien puisque l'on ne connaît pas les mécanismes biologiques d'entraînement du cerveau.

Réseaux de neurones récurrents et à mémoire (1982 puis 1993)

Ces RNN (Recurrent Neural Networks) permettent d'analyser des informations évoluant dans le temps comme la voix au niveau des phonèmes et le langage au niveau de l'assemblage des mots. Ils sont en effet très utilisés dans les systèmes de reconnaissance de la parole, pour la traduction automatique et la reconnaissance de l'écriture manuscrite.

Ils peuvent servir aux prévisions de cours d'action, de la consommation d'énergie ou d'eau pour les utilities, à l'analyse d'électrocardiogrammes (ECG)[125] et même à la détection des exoplanètes par la méthode des transits[126].

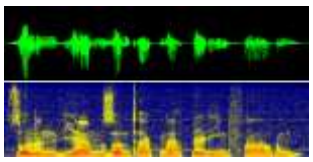
---

[124] Voir Learning representations by back-propagating errors, 1986.

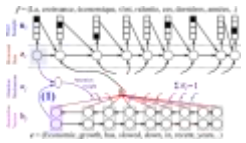
[125] Concommitamment avec d'autres méthodes comme les CNN, réseaux convolutifs.

[126] Que j'ai eu l'occasion d'expliquer ici : http://www.oezratty.net/wordpress/2017/astronomie-entrepreneuriat-exoplanetes/.

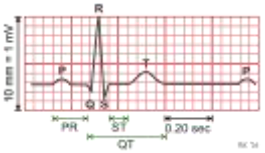
Dans l'industrie, on exploite aussi les RNN pour analyser le bruit ou les vibrations de machines pour y détecter des anomalies, dans le cadre de maintenance préventive. Il faut bien entendu en général disposer de données d'entraînement labellisées.



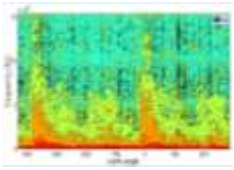
reconnaissance de la parole
(+ CNN)



traduction automatique

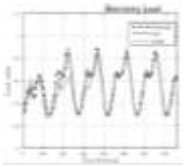


analyse d'ECG
(+CNN)



maintenance préventive

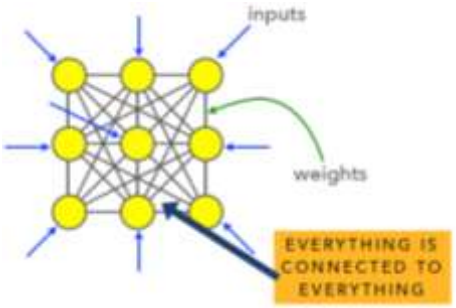
prévisions boursières



consommation

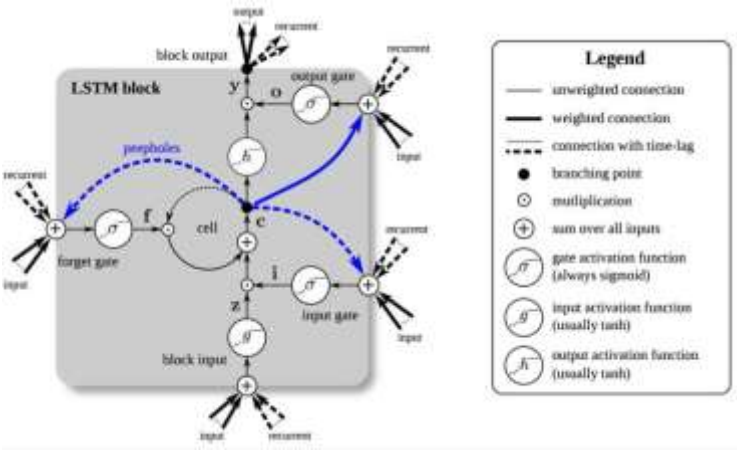
Les réseaux de neurones à mémoire ont bien évolué dans le temps avec différentes déclinaisons qui se sont inspirées les unes les autres :

- **Hopfield Network** (1982) est un des premiers réseaux à mémoire qui imite le fonctionnement de l'hippocampe du cerveau. Son créateur, un physicien, a relancé l'intérêt des réseaux de neurones qui étaient en rade depuis 1969. Le réseau de Hopfield permet de gérer une mémoire associative bidirectionnelle. Cela se retrouve dans sa matrice de connexions qui est symétrique et nulle dans sa diagonale. Tous les neurones du réseau sont connectés aux autres. La technique est cependant limitée en termes de stockage[127].



- **BPTT** (1987), BackProp Through Time, une méthode d'entraînement de réseaux de neurones récurrents, **RTRL** (1989), Real Time Recurrent Learning, une variante de réseaux de neurones récurrents puis **Simple Recurrent Network (SRN)** ou **Ellman Network** (1990) qui gère une couche cachée de contexte et s'entraîne par rétropropagation[128].

- **LSTM** (1997), Long Short Term Memory, qui savent gérer le contexte dans lequel les contenus apparaissent [129] et sont très utilisés pour le traitement du langage et la traduction automatique. Ce sont des réseaux en quelque sorte récursifs. Ils ont été créés par l'Allemand Sepp Hochreiter et le Suisse Jurgen Schmidhuber. Ils sont encore aujourd'hui une base clé du traitement du langage, sous forme de variantes diverses.



[127] Voir Hopfield Networks (29 slides).

[128] Voir The Simple Recurrent Network: A Simple Model that Captures the Structure in Sequences qui décrit bien la logique des SRN.

[129] Les LSTM ont été conceptualisés par Sepp Hochreiter et Jürgen Schmidhuber dans Long short-term memory, en 1997. Ce dernier est le créateur de la startup suisse **nNaisense**, qui ambitionne de créer une AGI (Artificial General Intelligence). Il est difficile d'obtenir des données de dimentionnement de ces réseaux de neurones. Les réseaux de neurones de traitement du langage de Google comprendraient 8 milliards de paramètres.

- **Reservoir Computing** (2005) utilise un « réservoir » de neurones constituant un système complexe et dynamique qui n'exploite pas l'approche classique du feed forward et qui pourrait s'apparenter à la notion de ligne à retard, comme dans l'audio[130]. On l'entraîne avec une fonction linéaire qui consolide les valeurs de l'ensemble du réservoir dans dans une couche de classification.

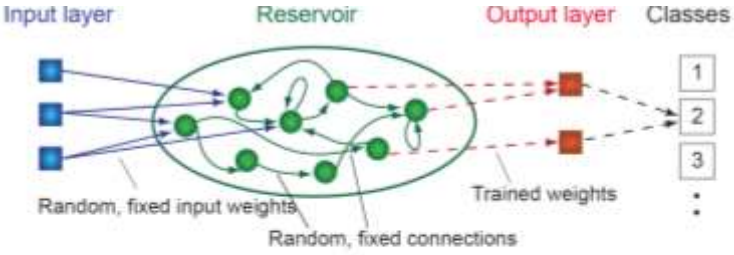


Fig. 1.8: Classical reservoir computing scheme. The input is coupled into the reservoir via a randomly connected input layer to the N nodes in the reservoir. The connections between reservoir nodes are randomly chosen and kept fixed, that is, the reservoir is left untrained. The reservoir' s transient dynamical response is read out by an output layer, which are linear weighted sums of the reservoir node states. Figure taken from Appeltant *et al.* [17].

Cette couche de classification ressemble à celle de la fin d'un réseau convolutif. Ce genre de réseau est adapté aux systèmes où règne le chaos. Il est utilisé dans les prévisions financières, dans le traitement du son, dans la détection d'épilepsies, dans la localisation de robots, dans la gestion de structures grammaticales. Il existe trois variantes de Reservoir Networks : Liquid State Machine, Echo State Network et Backpropagation-Decorrelation learning rule.

- **GRU** (2014), Gated Recurrent Units[131] est une des variantes plus simples des LSTM et est très utilisée. Elle évite notamment l'apparition du problème de diminution des gradients (« vanishing gradient problem ») qui empêche l'entraînement du réseau de neurones du fait de gradients de propagation d'erreurs trop faibles. Elle sert aussi à mémoriser les dépendances à longue distance dans les jeux de données.
- **BLSTM** (2015), Bidirectionnal Long Short Term Memory, des LSTM bidirectionnels, **Tree-LSTM** (2015), des LSTM organisés en arbres et non linéairement qui peuvent notamment servir à l'analyse de sentiment dans des textes[132], **Stacked RNN** (2015) qui sont des RNN empilés[133] puis **MANN** (2015), des Memory-Augmented Neural Networks qui permettent un entraînement plus rapide des réseaux à mémoire et avec un jeu de tests plus limité.
- **Transformers** (2017) sont des réseaux de neurones qui utilisent un mécanisme de gestion de l'attention[134]. Ils sont adaptés à la traduction et plus rapides à entraîner que les LSTM et autres variantes de réseaux récurrents[135]. Ils exploitent des méthodes d'apprentissage non supervisé. Le concept est issu de Google et notamment mis en œuvre dans GPT-3 qui est issu d'OpenAI en fait partie. Il prend la suite des réseaux récurrents et autres réseaux à mémoire (LSTM) qui tiennent compte de la temporalité des mots. Les transformers analysent le langage de manière parallélisée et non par séquences et en utilisant des mécanismes de gestion de l'attention. Cela explique pourquoi les nouveaux réseaux de neurones de type GPT-3 peuvent être entraînés sur de très grands jeux de données, 500 milliards de mots pour ce dernier, issus pour l'essentiel d'un crawl du web réalisé par l'association Common Crawl.

---

[130] Voir la présentation Introduction to Reservoir Computing, de Helmut Hauser, 2013 (282 slides) et la source du schéma Reservoir computing based on delay-dynamical systems (160 pages).

[131] Les GRU ont été créés par Junyoung Chung en 2014. Voir ce papier de Junyoung Chung, Caglar Gulcehre, KyungHyun Cho et Yoshua Bengio Empirical evaluation of gated recurrent neural networks on sequence modeling qui compare les GRU aux LSTM.

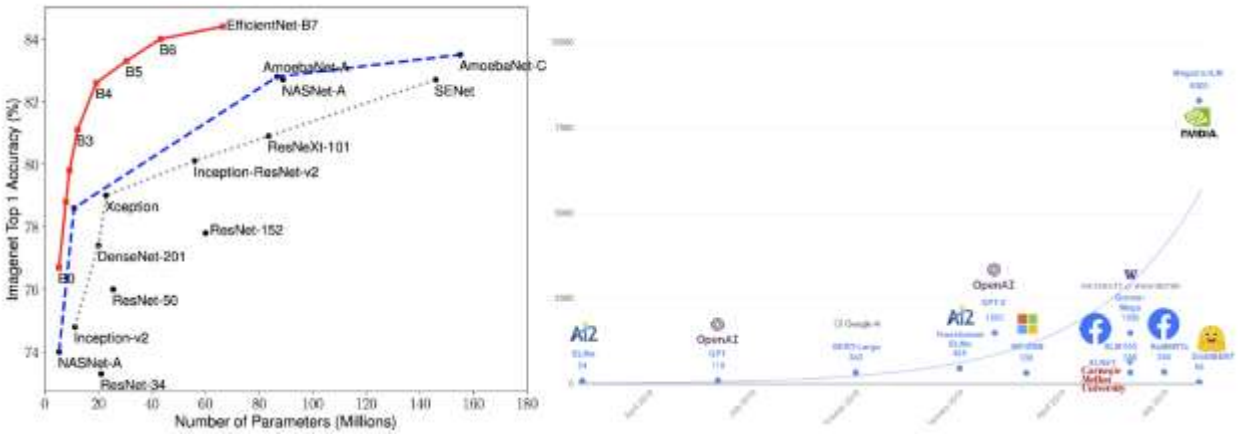
[132] Voir Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks de Kai Sheng Tai et Christopher Manning de Stanford et Richard Socher de Metamind, une startup acquise par Salesforce, 2015 (11 pages).

[133] Voir Meta-Learning with Memory-Augmented Neural Networks, 2016 (9 pages).

[134] Voir Decoding NLP Attention Mechanisms par Reda Affane, février 2020, qui explique bien le principe des transformers.

[135] Voir Attention Is All You Need, 2017 (11 pages), Universal Transformers, juillet 2018 (19 pages) et Transformer: A Novel Neural Network Architecture for Language Understanding de Jakob Uszkoreit, 2017.

Le tout alimentant un réseau de neurones comprenant 175 milliards de paramètres. Les transformers sont également utilisés dans la vision artificielle[136].



- **BERT** (2018, Bidirectional Encoder Representations from Transformers) est un nouveau modèle de réseaux de neurones originaire de Google qui donne d'excellents résultats dans la compréhension du langage[137]. Il a généré plein d'émules dont les Français **CamemBERT** et **FlauBERT**. Nous traitons de ces différents réseaux de traitement du langage dans une partie dédiée. Nous y évoquons notamment **GPT-3**.

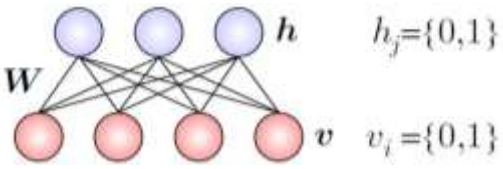
Je vous épargne les détails de toutes ces variantes de réseaux récurrents, ce d'autant plus que je n'ai pas encore très bien compris leur fonctionnement dans le détail. Ils sont difficiles à vulgariser, bien plus que les réseaux de neurones convolutifs que nous verrons un peu plus loin, et qui permettent d'analyser le contenu d'images[138].

Ces réseaux sont d'ailleurs souvent combinés entre eux de manière plus ou moins empirique. Ils transforment généralement les mots et phrases en vecteurs, des objets mathématiques triturés pour être comparés les uns aux autres, classifiés, modifiés et transformés. Ils permettent surtout de tenir compte du contexte dans lequel les objets comme des mots sont détectés pour analyser le sens d'une phrase. L'un des points clés de ces réseaux est leur capacité à mémoriser des contextes[139]. C'est un domaine d'amélioration encore plus intense que dans les réseaux de neurones convolutifs. Avec à la clé des solutions de plus en plus performantes pour la reconnaissance de la parole, la traduction automatique et les agents conversationnels réellement intelligents.

Machines de Boltzmann restreintes (1986)

Les machines de Boltzmann restreintes utilisent une seule couche de neurones source et cible.

Il n'y a pas de connexions entre les neurones d'une même couche. C'est le modèle le plus simple de réseau de neurones qui est ensuite exploité dans d'autres assemblages, comme les Deep Belief Networks (DBN) créés en 2006 par Geoff Hinton.



[136] Voir Essential Guide to Transformer Models in Machine Learning par Limarc Ambalina, octobre 2020, qui tente de vulgariser le fonctionnement des transformers.

[137] Voir BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding par Jacob Devlin, octobre 2018 (16 pages).

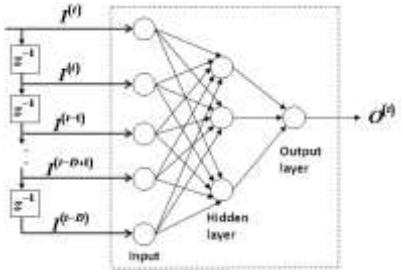
[138] Voir par exemple Understanding LSTM and its diagrams, par Shi Yan, 2016 qui explique comment sont assemblées les cellules des LSTM et le cours de Stanford Recurrent Neural Networks de Fei-Fei Li, Justin Johnson et Serena Yeung, mai 2018 (107 slides)

[139] Voir la conférence de Rob Fergus au Collège de France en avril 2016 dans le cadre de la chaire de Yann Le Cun.

Time Delay Neural Networks (1989)

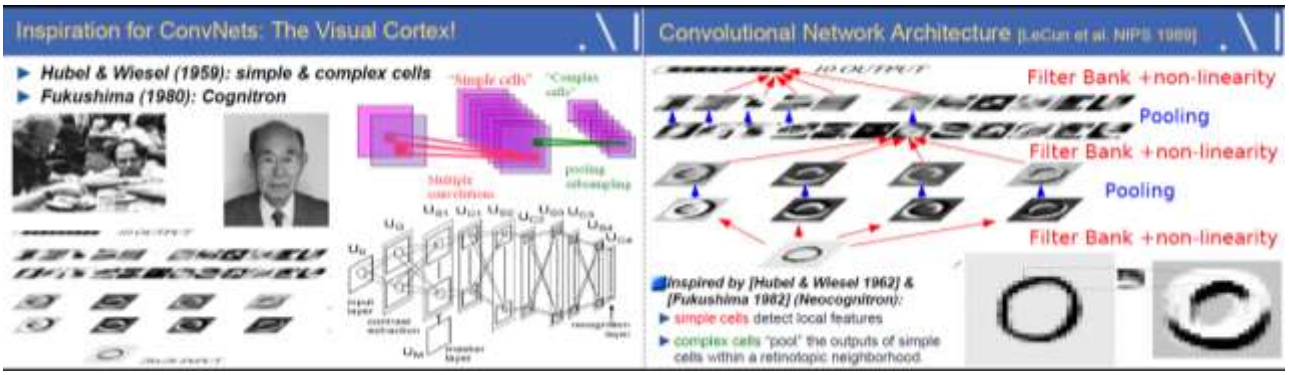
Créés par l'Allemand Alexander Weibel en 1989, les TDNN permettent notamment de reconnaître des phonèmes dans la parole sans avoir à les positionner explicitement dans le temps[140]. Les TDNN peuvent être considérés comme des précurseurs des réseaux convolutifs.

Ces derniers reconnaissaient au départ des images tandis que les TDNN reconnaissaient des sons qui sont aussi des données bi-dimensionnelle (fréquence, temps). La différence entre les TDNN et les ConvNets se situe surtout dans l'ajout de couches de pooling dans ces derniers. Nous verrons cela plus loin. Les TDNN ont été utilisés en France dans l'équipe de Françoise Soulié Fogelman pour la parole[141] et les images[142], et au LIMSI uniquement pour le traitement de la parole avec Laurence Devillers en 1992.



Réseaux de neurones convolutifs (1988)

En 1988, Yann Le Cun[143], qui avait quitté l'équipe de Françoise Soulié-Fogelman après sa thèse en 1987, pour rejoindre celle de Geoff Hinton à Toronto, puis celle de Larry Jackel aux Bells Labs en 1988, étend le modèle des TDNN à la reconnaissance de caractères puis à la reconnaissance d'images[144]. Cela a donné naissance aux réseaux de neurones convolutifs[145].



En 1993, on était capable de reconnaître un visage dans une image de 256x256 pixels, toujours sur une station de travail. Le temps de calcul d'une inférence était de 6 secondes sur une SPARCstation de **Sun Microsystems**. Les temps d'entraînement étaient aussi très longs et, qui plus est, ne pouvaient pas exploiter de gros volumes de données labellisées.

---

[140] Voir Review of TDNN (Time Delay Neural Network) – Architecture for Speech Recognition de Masahide Sugiyamat, Hidehumi Sazoait et Alexander Waibel qui font le point en 1991 de l'état de l'art des TNDD (4 pages).

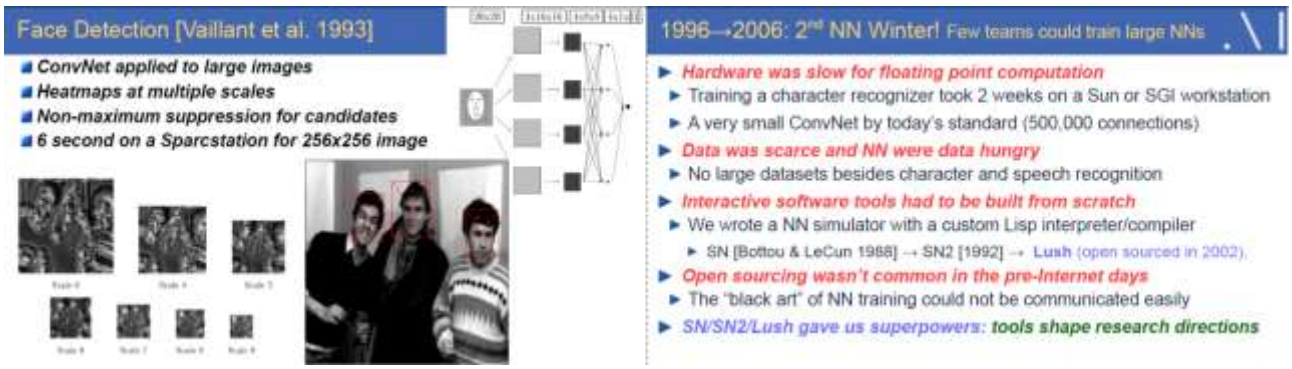
[141] Voir Experiments with time delay networks and dynamic time warping for speaker independent isolated digits recognition de Léon Bottou et al, 1989 (4 pages).

[142] Voir Scene segmentation using multiresolution analysis and MLP de Emmanuel Viennet et Françoise Fogelman Soulié, 1992.

[143] Yann Le Cun s'était inspiré des travaux de Kunihiko Fukushima, un chercheur de la NHK, et de ses réseaux de neurones multi-couches Neocognitron. Ce réseau de neurones pouvait reconnaître des chiffres mais manquait d'un algorithme d'apprentissage pour l'ensemble de ses couches. Voir Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition, 1987. Et le papier de 1998 Gradient-based learning applied to document recognition *dont les auteurs sont Yann Le Cun, Léon Bottou, Yoshua Bengio et Patrick Haffner (46 pages). L'illustration provient de la présentation* The Power and Limits Of Deep Learning, Yann Le Cun, mars 2019.

[144] Voir cette bonne explication en trois parties : A Beginner's Guide To Understanding Convolutional Neural Networks de Adit Deshpande (un étudiant aux USA), partie 1, partie 2 et partie 3, 2016.

[145] Yann LeCun raconte cela très bien dans Quand la machine apprend: La révolution des neurones artificiels et de l'apprentissage profond, octobre 2019.

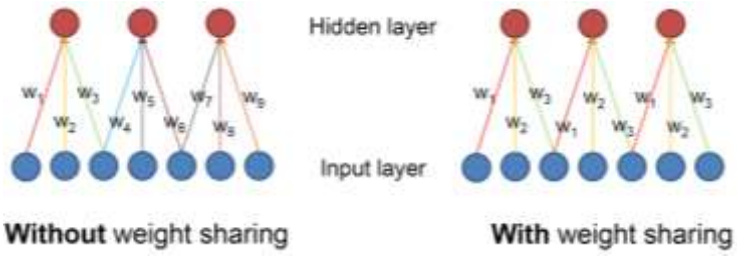


En 2012, le réseau **AlexNet** d'Alex Krizhevsky et Geoffrey Hinton écrasait les autres techniques lors de la compétition ILSVRC (ImageNet Large-Scale Visual Recognition Challenge), établissant la supériorité des réseaux de convolution pour la reconnaissance d'images face aux méthodes traditionnelles du machine learning. Ce sont des outils qui servent principalement à réaliser de la classification d'objets, comme pour associer une image à une classe d'objets (chat, bateau, avion…) ou un phonème vocal à son identifiant (/a/, /u/, ..).

Les premiers CNN de production ont été déployés en 1995 pour la reconnaissance des chèques, via une solution de NCR. Les CNN, appelés aussi ConvNets (convolutional neuron networks), utilisent plusieurs techniques enchaînées les unes avec les autres avec notamment des filtres et des feature maps qui consistent à identifier des formes dans les images, avec des niveaux d'abstraction allant du plus petit au plus grand. Depuis, la technique a évolué avec de nombreuses variantes de plus en plus performantes[146].

Un ConvNet utilise une alternance de convolution, d'activation non linéaire à base de ReLU puis d'agrégation, ou pooling. Les convolutions sont des transformations linéaires générant des **feature maps**, des matrices de pixels qui cartographient l'apparition d'un filtre donné dans l'image analysée. Un ConvNet utilise un jeu de plusieurs filtres initialisé aléatoirement, sauf pour la première couche de convolution qui est généralement initialisée avec des filtres classiques décrivant des transitions horizontales, verticales et diagonales dans les images, une technique utilisée dans les méthodes d'analyse d'images antérieures aux réseaux convolutifs. Les filtres sont des matrices de quelques pixels de côté, en général 3x3 ou 4x4[147]. Ils mettent en œuvre ce que l'on appelle le *weight sharing* (partage de poids) qui est illustré dans le schéma *ci-contre*.

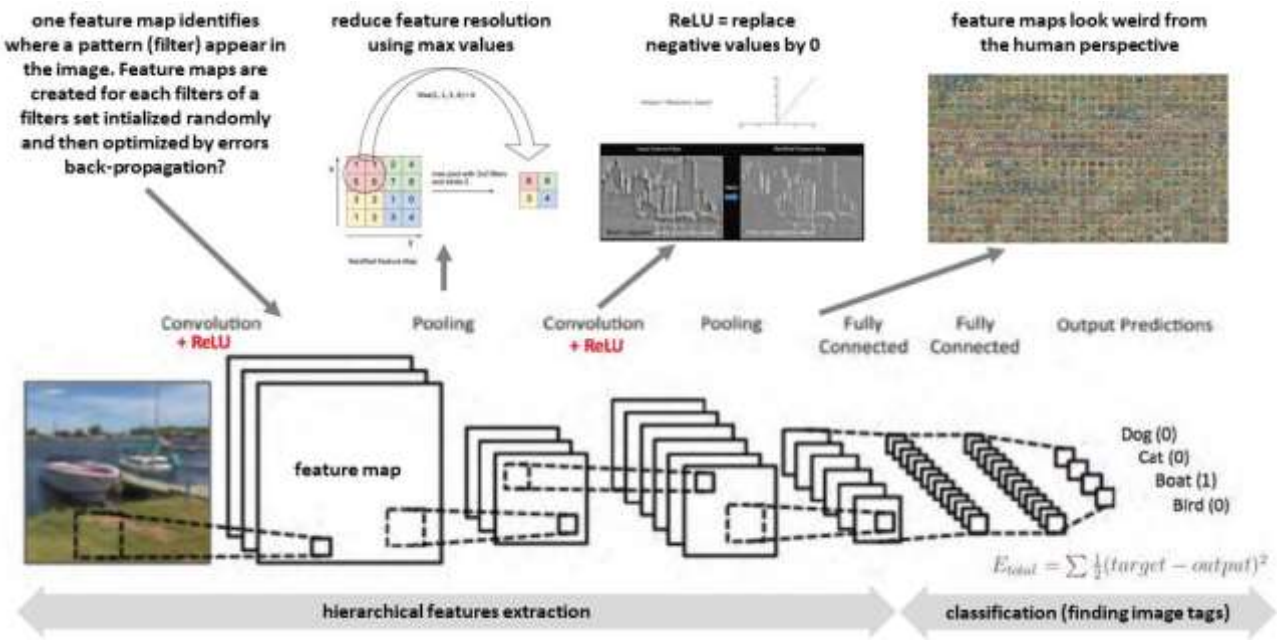
À gauche, un réseau fully connected où les neurones d'une couche sont reliés aux neurones d'une autre couche avec des poids différents. À droite, une convolution qui relie une couche de neurones à la suivante en partageant des poids qui correspondent aux filtres qui balayent l'image de la couche à traiter.



[146] Voir l'excellente série de papiers A Beginner's Guide To Understanding Convolutional Neural Networks, A Beginner's Guide To Understanding Convolutional Neural Networks Part 2 et The 9 Deep Learning Papers You Need To Know About (Understanding CNNs Part 3), 2016, qui décrit les différentes générations de CNN qui s'étaient alors succédées : AlexNet (2012, avec cinq couches de convolution, entraîné avec 15 millions d'images sur deux GPU Nvidia pendant une semaine et comprenant 60 millions de paramètres), ZF Net (2013), VGGNet (2014, qui utilise des petits filtres de 3x3 pixels), GoogleLeNet (2015, qui utilise des filtres de taille différente en parallèle et une centaine de couches en tout, devenu ensuite Inception dont la dernière version est la V4 sortie en 2016, puis Xception en 2018, voir Xception: Deep Learning with Depthwise Separable Convolutions par François Chollet, avril 2017, 8 pages), Microsoft ResNet (2015, avec 152 couches), les R-CNN (2013-2015), puis les GAN (2014). Voir aussi cette vidéo d'Andrej Karpathy, Deep Learning for Computer Vision, septembre 2016 (1h25)

[147] On retrouve cette taille de matrices dans les processeurs neuromorphiques et dans les derniers GPU de Nvidia Volta.

Ils sont ensuite affinés par rétropropagation d'erreurs de l'ensemble du réseau, un mécanisme qui est appliqué pour toutes les images d'un jeu d'entraînement qui peut comprendre des millions d'images[148] et même 3,5 milliards chez Facebook labellisées avec 15 000 hashtags différents[149]. Cet entraînement est très consommateur de ressources machine et aussi d'accès à la mémoire mais bien plus efficace qu'un simple réseau de neurones multicouches. Cela vient du fait que le réseau comprend moins de paramètres. Ce nombre de paramètres est approximativement égal à la somme de l'information des nombreux filtres de chaque convolution et des poids des synapses des couches terminales du réseau.



Chaque feature map générée par l'application des filtres sur l'image de départ se voit appliquée une suppression des valeurs négatives par une **ReLU** (Rectified Linear Units) pour séparer les contours bien nets, réduire la quantité de travail à appliquer aux couches suivantes et appliquer une transformation non linéaire. La ReLU est suivie d'une couche de **Max Pooling** qui applique une réduction de résolution permettant de réduire la sensibilité de la détection des formes à leur position dans l'image. Cette réduction se fait d'un facteur 2 dans les deux dimensions et conserve la valeur la plus élevée des petites matrices adjacentes de 2x2 pixels des feature maps réduites à un seul pixel.

Le processus est répété sur plusieurs couches, chaque feature map issue d'un niveau devenant une image qui subit un traitement équivalent dans la couche de convolution suivante. A la fin de l'histoire, la dernière couche de feature maps est reliée à une liste d'index de tags avec une probabilité de correspondance via quelques couches de neurones dites **fully connected**, à savoir que tous les neurones d'une couche sont liés à ceux de la couche suivante.

C'est là qu'un chat ou un bateau sont reconnus dans l'image. La dernière couche de cet empilement est un ensemble de neurones dont le nombre est égal au nombre de classes d'objets différents à reconnaître. Il peut être très grand mais doit rester raisonnable pour tenir compte des capacités du matériel. Plus est grand le nombre d'objets différents que l'on veut reconnaître, plus devra être grand le nombre de paramètre du réseau de neurones.

Un ConvNet de type **VGG16** à cinq couches de convolution et prenant en entrée des images de 224x224 pixels en couleur (RGB) contient 138 millions de paramètres.

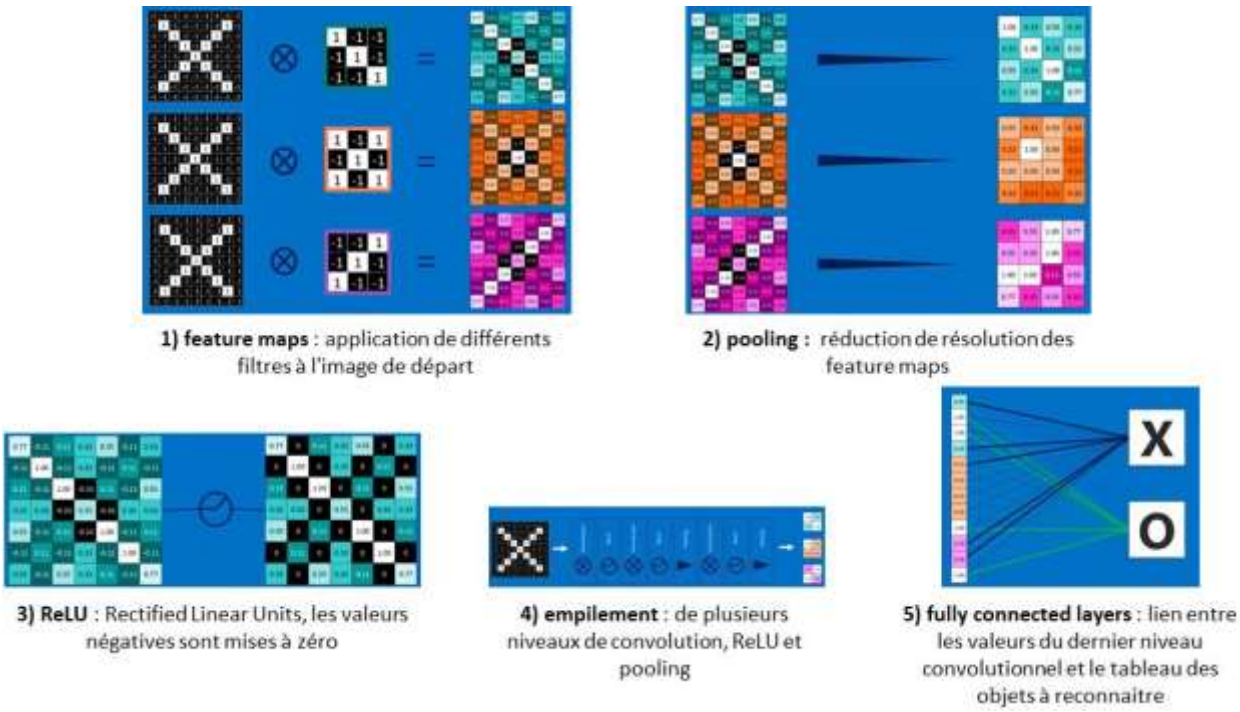
---

[148] Voir Everything you wanted to know about Deep Learning for Computer Vision but were afraid to ask (25 pages) qui décrit bien les techniques d'optimisation de la phase d'entraînement d'un réseau convolutionnel.

[149] Voir Advancing state-of-the-art image recognition with deep learning on hashtags, de Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Manohar Paluri et Laurens Van Der Maaten, mai 2018.

Pour les images en couleur, les filtres de la première couche de convolution sont des matrices nxnx3, n étant la taille des filtres. La feature map résultante est une matrice carrée avec une seule couche de profondeur. Chaque point de cette matrice est le résultat de la comparaison entre le filtre à trois couleurs avec des matrices de même taille extraites de l'image d'origine. La couleur peut être représentée en mode RGB ou dans un espace colorimétrique YUV, avec une luminance, la saturation et la chrominance.

Les moteurs de reconnaissance d'images reconnaissent au grand maximum que quelques dizaines de milliers de classes d'objets dans cette dernière couche de réseaux de neurones[150]. C'est lié en particulier aux contraintes de la mémoire des GPU utilisés. Les derniers GPU Nvidia qui sont couramment utilisés pour entraîner des ConvNets disposent de 16 Go à 32 Go de mémoire.



Voici, *ci-dessus*, un autre exemple illustré du processus des ConvNet de reconnaissance de caractères. En 1), on peut identifier la présence des diagonales et croix dans les feature maps à droite.

Puis le pooling en 2) pour divise par deux la résolution des feature maps, la couche ReLU qui fait un max(0, x) sur toutes les valeurs (avant ou après le pooling), puis en 5), les couches de neurones qui aboutissent au résultat final indiquant la valeur de la lettre. Selon les modèles, des variantes diverses sont introduites dans ces couches qui visent en général à augmenter le contraste de l'image traitée.

A chaque niveau d'un réseau convolutif, le nombre de feature maps augmente et leur taille diminue. Les feature maps étant optimisées automatiquement, leur forme n'est pas vraiment interprétable par le cerveau humain, alors que les filtres du traitement classique d'images antérieur aux réseaux convolutifs l'étaient.

C'est la magie des ConvNets : ils créent des niveaux de représentations hiérarchiques intermédiaires des images qui optimisent leur reconnaissance, sans que l'on puisse comprendre comment ils fonctionnent pas à pas et dans le détail[151].

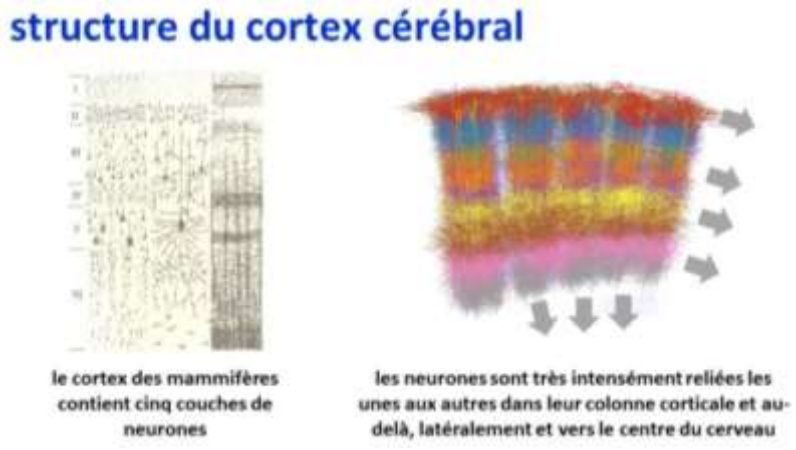
---

[150] D'ailleurs, certaines démonstrations étonnantes de reconnaissance d'objets oublient de préciser le nombre d'objets que le système peut reconnaître !

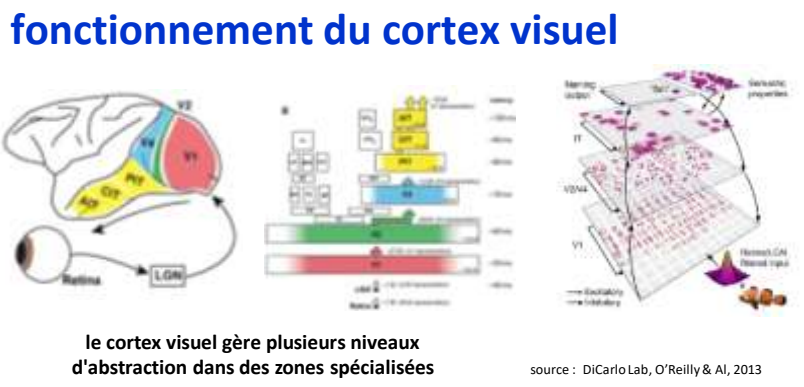
[151] Par contre, comme la résolution des feature maps diminue de couche en couche, la détection de l'emplacement des macro-objets détectés est très mauvaise. Elle est même quasiment inexistante.

D'où la fameuse « non explicabilité » des algorithmes qui inquiète nombre d'observateurs[152], ce d'autant plus qu'elle se produit aussi dans les réseaux récurrents et à mémoire qui servent principalement au traitement du langage. Mais, pour beaucoup d'applications, ce qui compte avant tout est la qualité des résultats plus que leur explicabilité. En cas de défaillance d'un réseau de neurones, l'erreur proviendra probablement d'une base d'entraînement ne couvrant pas bien l'espace des possibilités que le réseau peut rencontrer dans sa mise en production.

Nous en reparlerons plus loin au sujet du biais des données d'entraînement. Les ConvNets s'inspirent fortement du mode de fonctionnement du cortex visuel des mammifères qui est structuré, de près, dans des colonnes corticales faites de cinq couches de neurones et qui, de loin, comprend des aires spécialisées qui élèvent progressivement le niveau d'abstraction des objets reconnus (*ci-contre*) [153].



On retrouve cette architecture à cinq couches dans bon nombre de réseaux de neurones convolutifs, sachant que ces couches peuvent elles-mêmes comprendre de nombreuses sous-couches de neurones. Par contre, contrairement au cortex humain, les ConvNets qui font de la reconnaissance d'images utilisent des représentations à très basse résolution.



La majorité des ConvNets se contentent d'images comprimées à une résolution voisine de 224x224 pixels et rarement au-delà de 256 pixels. Il a fallu que je décortique le convnet **AlexaNet** qui date de 2012 pour mieux comprendre la manière dont les couches de convolution étaient reliées entre elles. Les schémas qui décrivent ce genre de convnet peuvent facilement donner le mal de tête[154]. J'ai eu mal à la tête mais j'ai enfin compris ! En voici donc une explication.

Le principe décrit visuellement *ci-dessous*, consiste à utiliser des filtres multidimensionnels. Dans les cinq couches de convolution d'AlexNet, les filtres sont des matrices de respectivement 11x11x3 pixels pour la première couche de convolution (11 pixels de côté x 3 couleurs), 5x5x96 pixels pour la seconde (96 étant le nombre de filtres de la première couche de convolution et de feature maps qui en résultent), 3x3x256 pixels pour la troisième (256 étant le nombre de filtres et feature maps de la seconde couche de convolution), 3x3x384 pixels pour la quatrième (384 étant le nombre de filtres de la troisième couche de convolution) et 3x3x384 pixels pour la cinquième (bis repetita pour la quatrième couche de convolution).

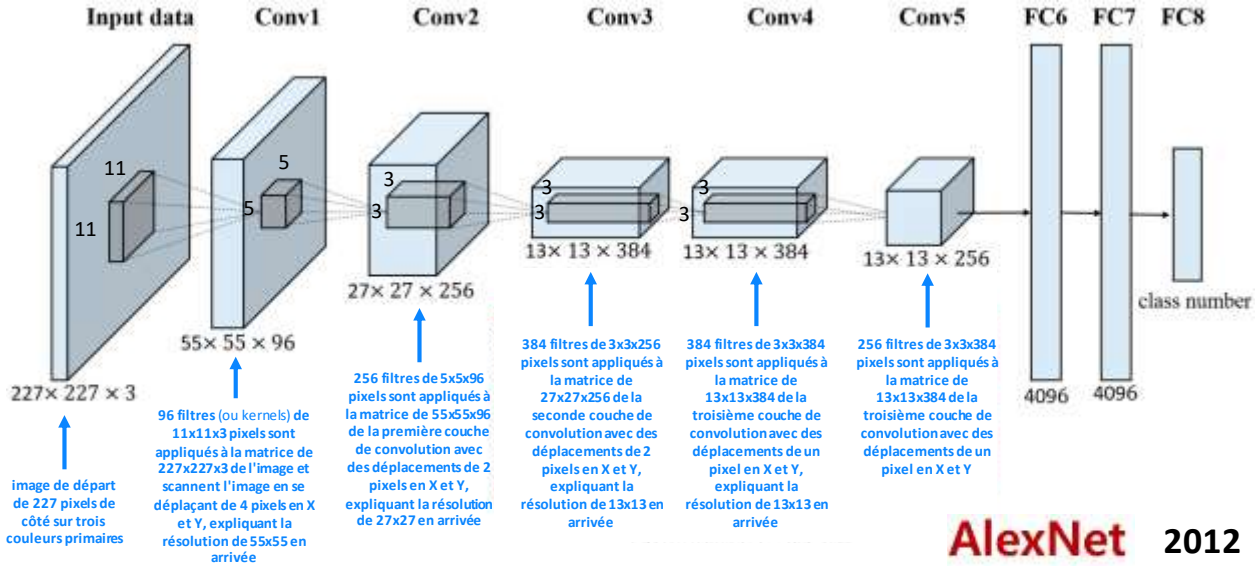
---

[152] Voir Le talon d'Achille de l'intelligence artificielle de Benoit Georges, mai 2017.

[153] Voir Receptive fields, binocular interaction and functional architecture in the cat's visual cortex de David Hubel et Torsten Wiesel, 1962 ainsi que l'ouvrage de référence Eye, Brain, and Vision par David Hubel, 1995 (240 pages) qui décrit dans le détail le fonctionnement du système de vision étudié au départ à partir de celui du chat. La première couche V1 détecte l'orientation des formes par angles de 6°.

[154] Voir ImageNet Classification with Deep Convolutional Neural Networks d'Alex Krizhevsky, Ilya Sutskever et Geoffrey Hinton de l'Université de Toronto, 2012 (9 pages).

Ces filtres à trois dimensions sont dénommés, en anglais, des *multiple input channels*[155].



Comment passe-t-on de 227 à 55 puis à 13 pixels de côté entre l'image de départ et les couches de convolution suivantes ? Cela vient du fait que les filtres de la première convolution se déplacent de 4 pixels en 4 pixels (avec du recouvrement puisqu'ils font 11 pixels de côté), ceux de la seconde et de la troisième se déplacent de 2 pixels en 2 pixels (avec également du recouvrement puisqu'ils font 5 et 3 pixels de côté) et les deux derniers de 1 en 1 pixel. Ce déplacement est dénommé « stride » en anglais. Cette méthode est l'équivalent d'une réduction de résolution avec une couche de pooling mais elle semble plus efficace et rapide qu'un scan pixel par pixel des images dans chaque convolution puis une réduction de résolution d'un facteur 4 ou 2.

Mathématiquement, l'application d'un filtre à un morceau d'image ou de couche de convolution de la même taille consiste à multiplier les valeurs des pixels du filtre à celle de l'image ou de la couche de convolution scannée, à additionner le tout, à ajouter un éventuel coefficient de correction appelé un biais et à y appliquer une fonction d'activation qui va normaliser le résultat entre 0 et 1. C'est souvent une fonction sigmoïde. Cette opération mathématique va permettre d'identifier la dose de points communs entre le filtre et le bout de convolution analysé qui fait la même taille que le filtre. L'objet détecté est très « macro » puisque le filtre est multidimensionnel.

La dernière convolution d'AlexaNet génère une matrice de 13x13x256 pixels qui alimente à son tour trois couches de neurones « fully connected » permettant d'identifier 1000 objets différents. La description d'AlexNet dans l'article d'origine est encore plus tarabiscotée, *ci-dessus*. Cela vient du fait que les convolutions sont découpées en deux parties (en haut et en bas) pour être réparties sur deux GPU. C'était l'un des premiers réseaux convolutifs avec un entraînement distribué sur plusieurs processeurs. Aujourd'hui, on arrive à le faire sur un très grand nombre de GPUs.

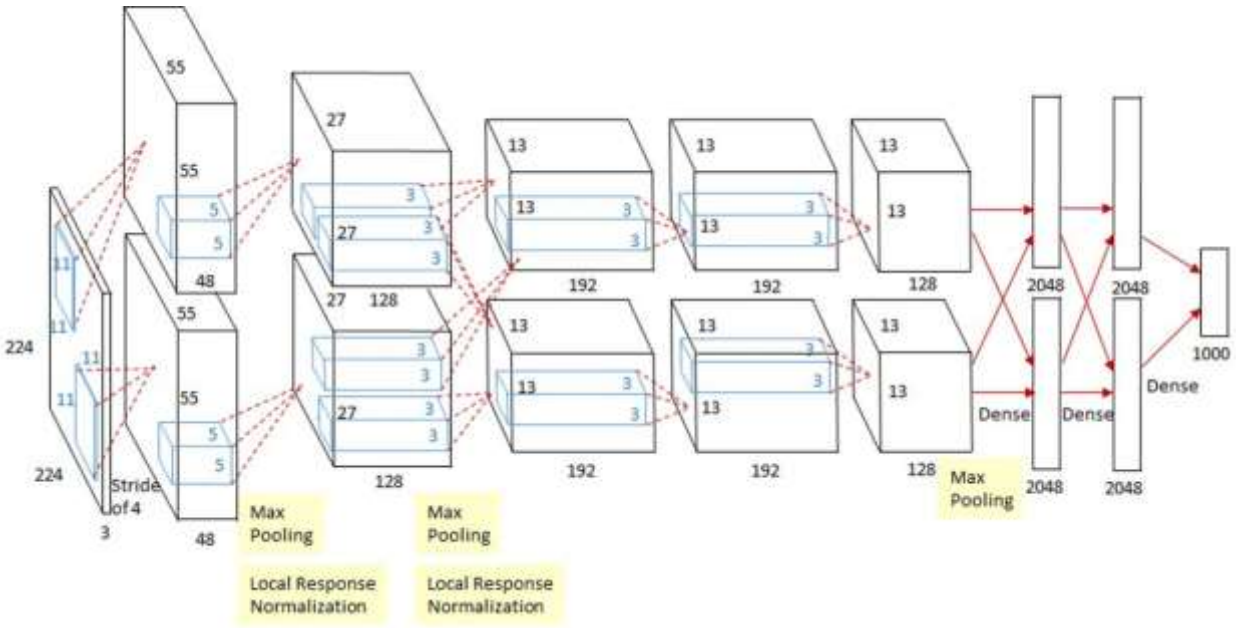
Créés en 2015, les **ResNet-50** utilisent 48 couches de convolution et deux de pooling avec un système de court-circuits de couches permettant de faire converger le réseau à l'entraînement.

Les CNNs ont aussi progressé pour permettre la détection puis la segmentation fine de plusieurs objets dans une image. Le premier modèle adapté était le **R-CNN** (Regions with CNN, 2013[156]) qui était capable de détecter chacun des objets de l'image, de les labelliser, et de fournir les coordonnées du cadre les contenant. Il est entraîné par des images d'objets déjà « croppés ».

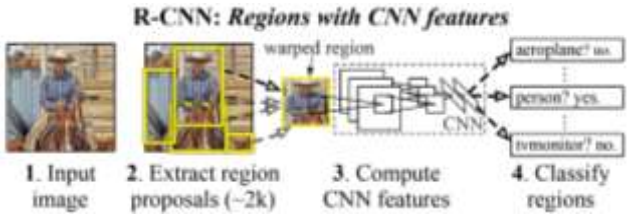
---

[155] On trouve une bonne vulgarisation de ces composantes des réseaux convolutifs dans Convolutional Neural Networks par Jia-Bin Huang, Virginia Tech, 2016 (88 slides).

[156] Voir Rich feature hierarchies for accurate object detection and semantic segmentation par Ross Girshick & Al, de l'Université de Berkeley, novembre 2013 (21 pages).



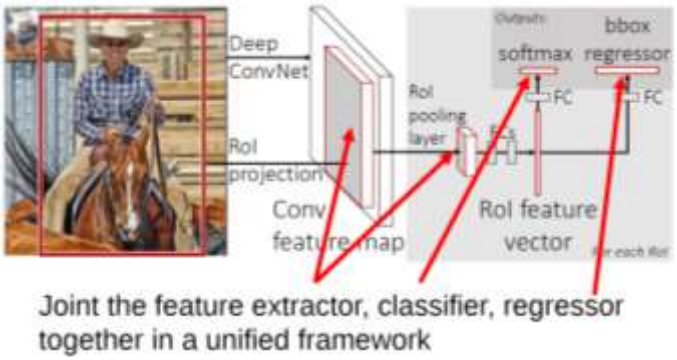
Il recherche ces cadres avec une méthode de « recherche sélective » qui teste environ 2000 cadres déterminés en fonction des transitions dans les images. Les images « croppées » sont ensuite injectées dans une version d'AlexNet pour labelliser l'objet qu'il contient.



Un discriminateur à base de SVM permet ensuite de « scorer » la reconnaissance d'objet puis de les labelliser. Le modèle utilise enfin une régression linéaire pour déterminer le cadre qui s'approche le plus de l'objet. Le procédé était très lourd car nécessitant de faire passer plusieurs milliers d'extraits de l'image dans un CNN. Qui plus est, les cadres devaient avoir une forme fixe, les images étaient donc déformées pour générer un carré avant le passage dans le CNN de reconnaissance et de labellisation.

Ce procédé s'est d'abord perfectionné avec **SPP-Net** qui analyse l'image à différents niveaux de résolution, de manière « pyramidale », permettant de réduire d'un facteur 24 à 64 la quantité de calculs dans l'entraînement et aussi de détecter des objets quelle que soit leur taille[157].

Suivit le **Fast R-CNN** (2015[158]) qui évite une partie du passage répétitif des images croppées dans un CNN avec un processus de détection des régions d'intérêt qui est réalisé après l'application d'une première couche de convolution, exploitant une technique appelée « Region of Interest Pooling ». Ce procédé accélère les traitements d'un facteur 10 par rapport aux SPP-Net et améliore au passage son efficacité.



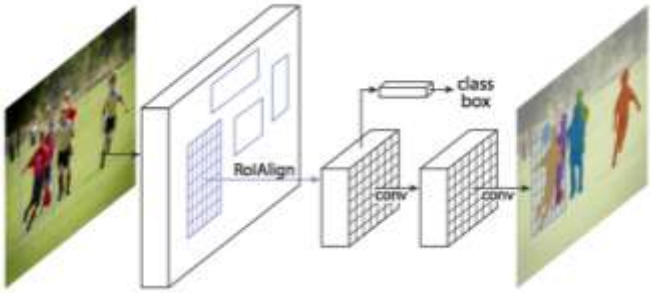
Le Fast R-CNN est aussi bien plus intégré, consolidant le CNN, le classifieur et la régression de détection de cadre d'objets dans un modèle unique.

---

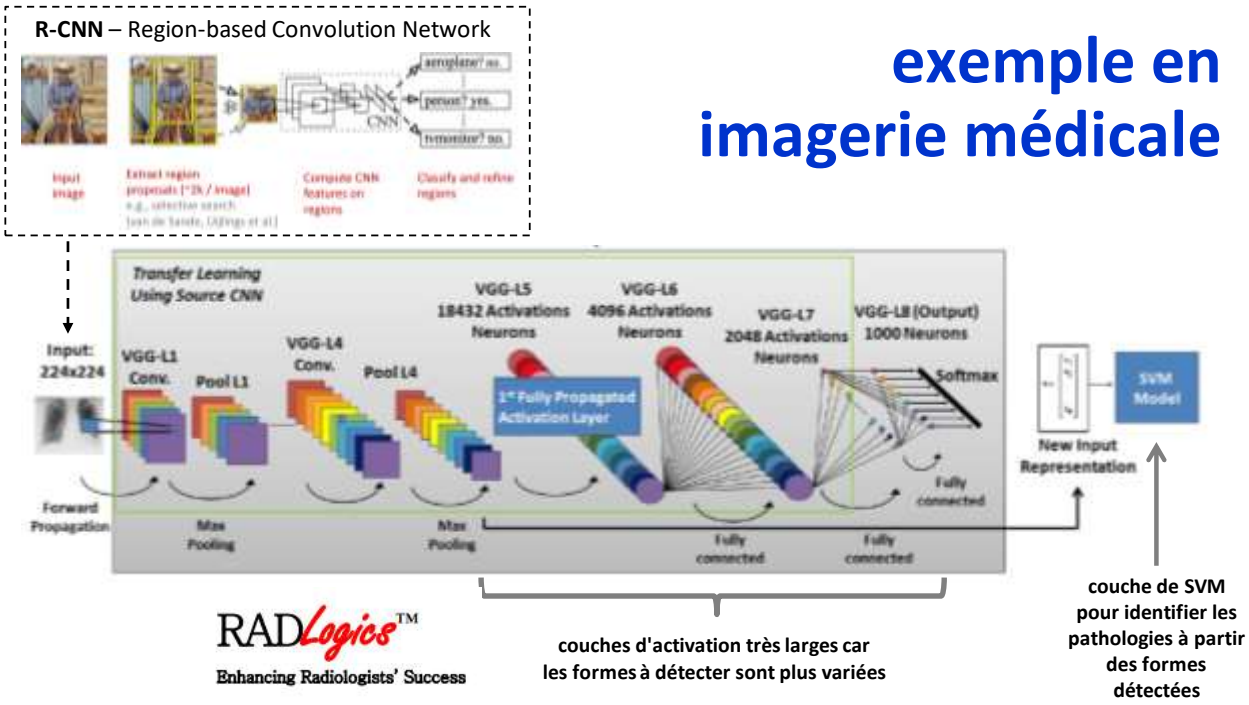
[157] Voir Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition par Kaiming He & Al, juin 2014 (14 pages). Voir aussi Feature Pyramid Networks for Object Detection, 2016.

[158] Voir Fast R-CNN par Ross Girshick de Microsoft Research, avril 2015 (9 pages).

Se sont alors succédés **Faster R-CNN**[159] qui accélère la détection des cadres des objets puis enfin **Mask CNN**[160], originaire de Facebook, qui étend le Faster R-CNN pour segmenter précisément les objets au niveau des pixels, y compris lorsqu'un objet est occulté par un autre (exemples *ci-dessous à droite*).



Ces réseaux peuvent être combinés dans le cadre de l'imagerie médicale qui est plus exigeante. Ainsi, dans cet exemple composite de détection, labellisation et décompte de nodules cancéreux dans une radio du poumon va-t-on d'abord utiliser un R-CNN ou équivalent qui va d'abord décompter les nodules cancéreux de manière générique avec une seule classe d'objet détectée et dans une version 224x224 pixels de la radio. Il est probable que cette combinaison de réseaux de neurones doit pouvoir s'intégrer dans un seul réseau.



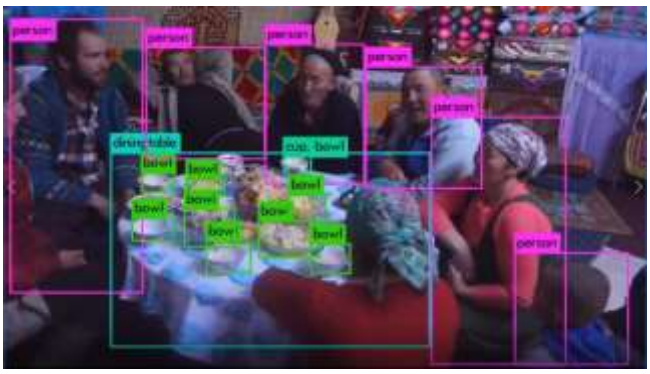
Les coordonnées de chaque nodule vont ensuite être exploitées pour les extraire dans l'image source haute résolution de la radio. Cette version plus haute définition, qui pourra atteindre 224x224 pixels - et éventuellement détourée - de chaque nodule sera alors injectée dans un CNN classique de classification qui permettra de labelliser précisément le type de nodule dont il s'agit.

---

[159] Voir Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks par Shaoqing Ren & Al, janvier 2015 (14 pages).

[160] Voir Mask R-CNN par Kaiming He & Al, mars 2017 (12 pages). Voir cet historique et explications associées : A Brief History of CNNs in Image Segmentation: From R-CNN to Mask R-CNN de Dhruv Parthasarathy, avril 2017.

Au passage, les images des radios sont reconnues en couleur sauf lorsqu'une version monochrome de l'image est suffisante pour obtenir une reconnaissance avec un faible taux d'erreur, ce qui est souvent le cas dans l'imagerie médicale qui est habituellement réalisée en noir et blanc (radios, IRM, échographies). Il existe parfois des images réalisées en fausses couleurs pour représenter des niveaux de densité différents, ou lors de l'intégration d'images provenant de systèmes différents (IRM et scanner rayons X par exemple).



La détection d'objets multiples dans une image peut faire appel à **Yolo V3** (2016) ou « You only look once », qui fait de la détection d'objets multiples dans des images, notamment dans les vidéos. Il fonctionne en temps réel et jusqu'à 45 images par secondes, selon le matériel utilisé pour les inférences. Yolo V3 utilise 24 couches de convolution et deux couches « fully connected » en sortie avec des entrées en 224x224 pixels. La technique est voisine de celle des Fast R-CNN[161].

Pourquoi donc 224 pixels de côté dans les CNN, R-CNN et autres Yolo ? Parce qu'on peut diviser cette valeur plusieurs fois par 2 et aboutir à 7, qui est la taille des feature maps de la dernière couche de convolution. Malgré la légère diminution de résolution entre une image et sa feature map, celle-ci reste stable. Cela vient du fait que l'on ajoute un pixel à 0 sur les bords de l'image avant de la balayer avec les filtres qui font souvent 3 pixels de côté. La feature map de la première couche fait ainsi 224 + 2 (pixels de bord) – 3 (taille des filtres) +1 (effet de bord) = 224 pixels. Les couches suivantes font 112, 56, 28, 14 puis 7 pixels. Et voilà l'explication !

Mais il peut y en avoir des variantes avec des nombres de pixels avoisinants et des feature maps de taille variable selon les couches comme dans AlexNet (avec des filtres de 11, 5 et 3 pixels de côté) comme nous l'avons vu précédemment.

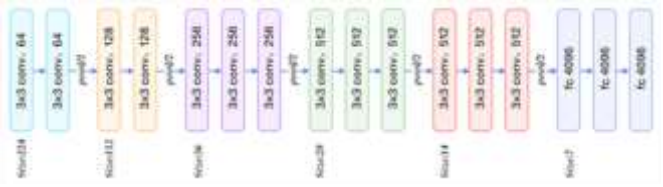
Afin de réduire la charge de calcul et de réduire la consommation d'énergie, les valeurs gérées en entrée et en sortie de neurones sont souvent encodées dans des entiers sur 4 ou 8 bits au lieu de nombres flottants 16 ou 32 bits. C'est pour cela que la performance des processeurs embarqués est souvent décrite en GigaOps et pas en GigaFlops. Les « Ops » sont des opérations sur des entiers (4, 8 ou 16 bits à préciser) et les « Flops » sur des nombres flottants. Un GigaOps n'est donc pas équivalent à un GigaFlops !

Les algorithmes utilisés sont cependant si puissants qu'ils permettent de générer des taux de reconnaissance d'images meilleurs que ceux de l'Homme ! Qu'est-ce que cela serait si la résolution utilisée était la même que dans l'œil et le cortex humains, l'œil étant doté de 90 millions de bâtonnets pour détecter la luminance et de 6,5 millions de cônes pour la couleur, le tout étant connecté au cortex visuel par un nerf optique comprenant un million d'axones !

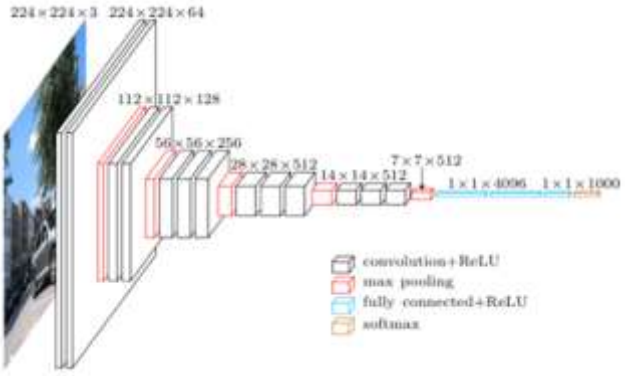
On peut distinguer les ConvNets selon le nombre de dimensions des données reconnues : 1D (une dimension) pour le texte, la reconnaissance de genre de musique, des prévisions temporelles sur une seule variable, 2D (deux dimensions) pour les images, pour la reconnaissance de la parole qui associe fréquence audio et temps, puis 3D (trois dimensions) pour le traitement de vidéos et d'imagerie médicale 3D.

[161] Voir You Only Look Once: Unified, Real-Time Object Detection, (10 pages) et cette vidéo. Le projet est open source. Le créateur de Yolo est Joe Redmond, le fondateur de la startup Xnor.ai (2016, USA, $14,6M) qui package des solutions logicielles de deep learning pour des besoins courants comme la reconnaissance d'objets dans des images. En 2020 sont sortis presque simultanément les versions 4 et 5 de Yolo, issues de différents auteurs et qui améliorent toutes les deux les performances de l'inférence par rapport à la V3.

Il existe un très grand nombre d'algorithmes de ConvNets[162]. L'un des plus courants est **VGG16** (2014) qui comprend 16 couches de convolution avec des filtres de 3x3 pixels, de 64 à 512 filtres et features maps par couche et la capacité de détection de 1000 classes d'objets différentes.

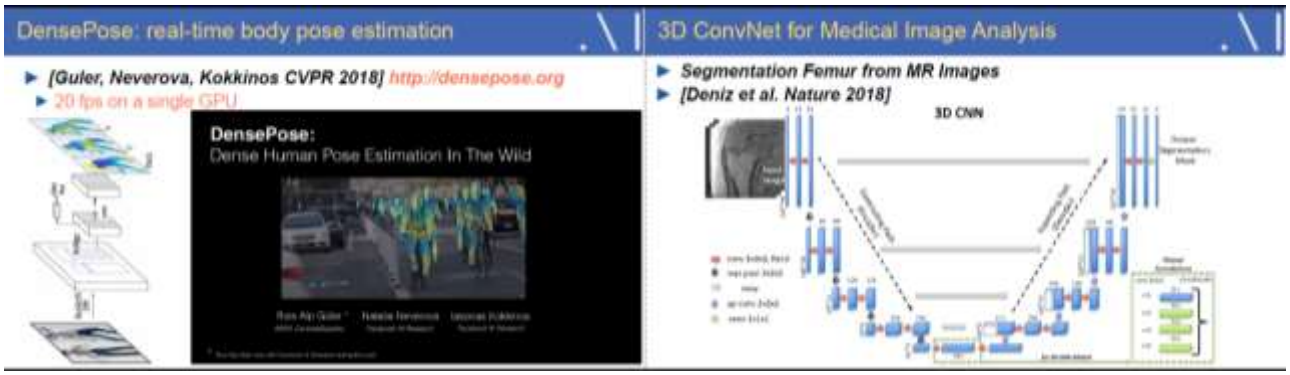


Le modèle est alimenté par 528 Mo de paramètres. Plus on va augmenter le nombre de couches et le nombre de filtres par couches, plus on va augmenter la taille mémoire nécessaire pour la gestion de ces paramètres, de type HBM2 dans les Nvidia V100. Cela explique pourquoi les Convnets analysent des images à assez basse résolution pour le moment, même si certains peuvent aller au-delà de 227x227 pixels avec l'ajout de grandes quantités de mémoire dans les serveurs d'entraînement.



Une équipe du MIT créait en 2018 l'algorithme "Proxyless neural architecture search" qui permet d'accélérer les optimisations de modèles d'IA, notamment dans la classification d'images[163]. Il passe par la réduction des besoins en mémoire de GPU pour l'entraînement, par la suppression de branches inutiles des réseaux de neurones pour les simplifier, puis par l'intégration des temps de latence des composants matériels exploités. Ils utilisent aussi des filtres 7x7 au lieu de 3x3 ou 5x5[164].

Citons aussi deux exemples d'évolutions de ces méthodes. **DensePose** (2018) permet d'identifier la posture d'humains à un rythme de 20 images par secondes sur un seul GPU Nvidia GTX1080 (vidéo). Il exploite un R-CNN[165]. C'est un projet du FAIR de Facebook. Un autre exemple de réseau de neurones convolutif 3D est exploité pour l'analyse d'IRM des os[166] (2017).



[162] Voir ce cours de Stanford sur les CNN : CNN Architectures de Fei-Fei Li, Justin Johnson et Serena Yeung, mai 2018 (106 slides).

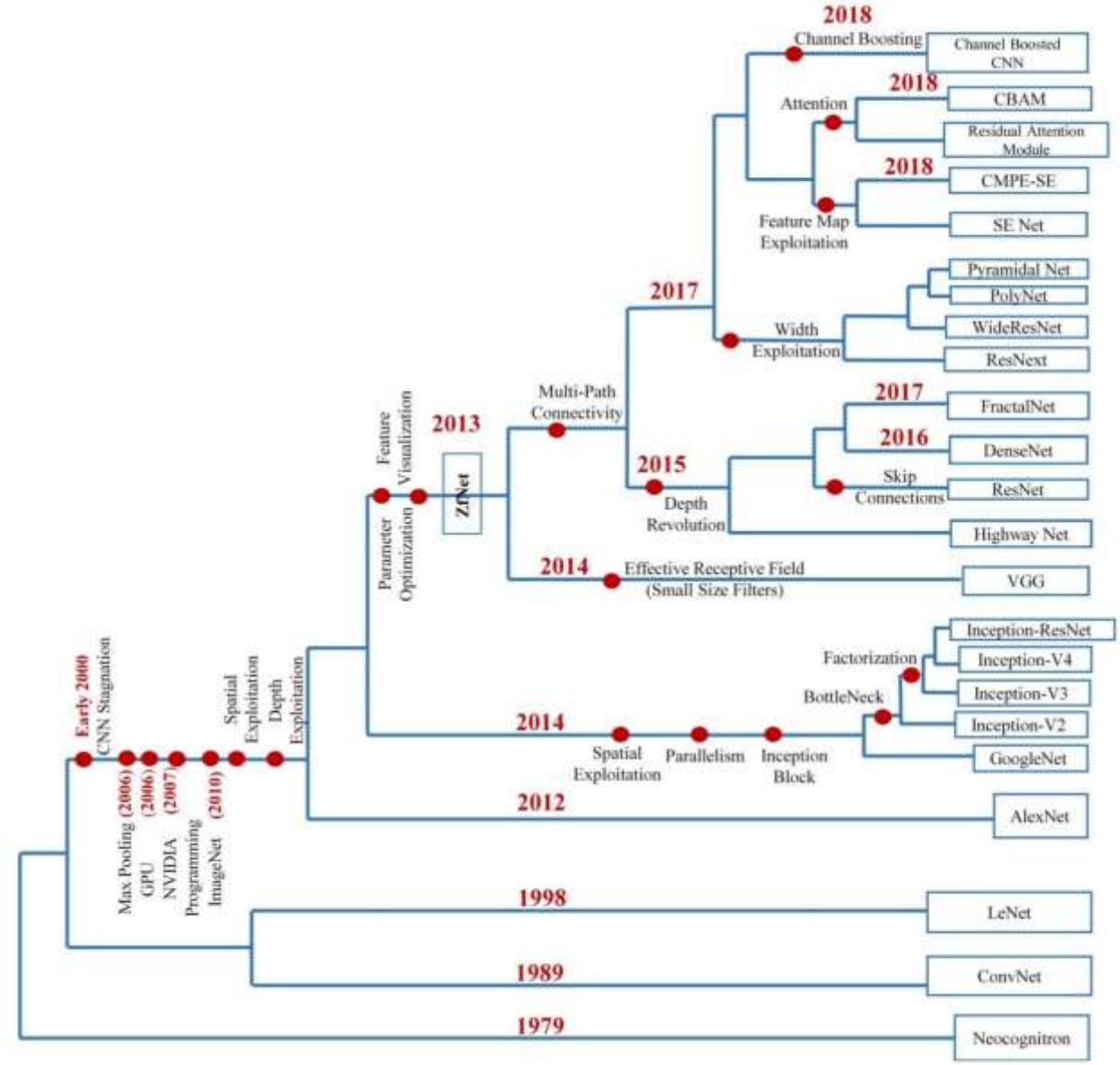
[163] Voir ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware par Han Cai & Al, décembre 2018.

[164] Voir Using AI to Make Better AI par Mark Anderson, avril 2019, qui fait référence à ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware par Han Cai & al, décembre 2018 (13 pages).

[165] Voir DensePose: Dense Human Pose Estimation In The Wild, Natalia Neverova & Al, février 2018 (12 pages). L'un des contributeurs du papier est Rıza Alp Guler d'Inria et CentraleSupelec.

[166] La source est toujours The Power and Limits Of Deep Learning, Yann Le Cun, mars 2019. Voir Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Networks, Cem Deniz & Al, Université de New York, 2017 et révisé en 2019 (13 pages).

Une étude de 2019 dresse un beau panorama historique et technique de l'évolution des réseaux convolutifs entre 1979 et 2018, avec cette généalogie intéressante[167]. L'une des évolutions les plus récentes concerne la recherche de frugalité des convnets, notamment pour leurs phases d'entraînement. Nous l'évoquons dans une partie de cet ouvrage dédiée aux questions d'énergie.



Apprentissage par transfert (1997)

L'apprentissage par transfert permet de réaliser de l'apprentissage incrémental dans les réseaux de neurones, aussi bien convolutifs et à mémoire. Il permet de transférer ce qui a été appris par un réseau dans un autre réseau, et même dans un domaine un peu différent[168].

Par exemple un réseau ayant appris à reconnaître des chats sert à initialiser un réseau pour reconnaître des chiens. Cette possibilité provient de ce que les primitives de bas niveau, les filtres des premières couches, sont les mêmes pour des familles d'images du même genre.

---

[167] Voir A Survey of the Recent Architectures of Deep Convolutional Neural Networks par Asifullah Khan et al, 2019 (70 pages).

[168] Voir Transfer Learning de Lisa Torrey et Jude Shavlik (22 pages, non daté), Transfer Learning : réaliser de meilleures prédictions avec peu de données d'Arthur Letang, janvier 2017, A Gentle Introduction to Transfer Learning for Deep Learning de Jason Brownless, décembre 2017, Understanding, generalisation, and transfer learning in deep neural networks, février 2017, et les présentations Transfer Learning (38 slides) et Recent Advances in Transfer Learning, 2018 (28 slides). Et enfin Voir Deep learning isn't hard anymore par Caleb Kaiser, 2020.

C'est une discipline à part entière du deep learning qui a de nombreuses variantes et applications, notamment pour réduire la quantité de données nécessaires à l'apprentissage d'un réseau de neurones.

La méthode permet aussi d'entraîner un réseau de neurones de manière incrémentale, lorsque l'on ajoute par exemple des images à reconnaître dans un convnet déjà entraîné avec un jeu d'images initial[169].

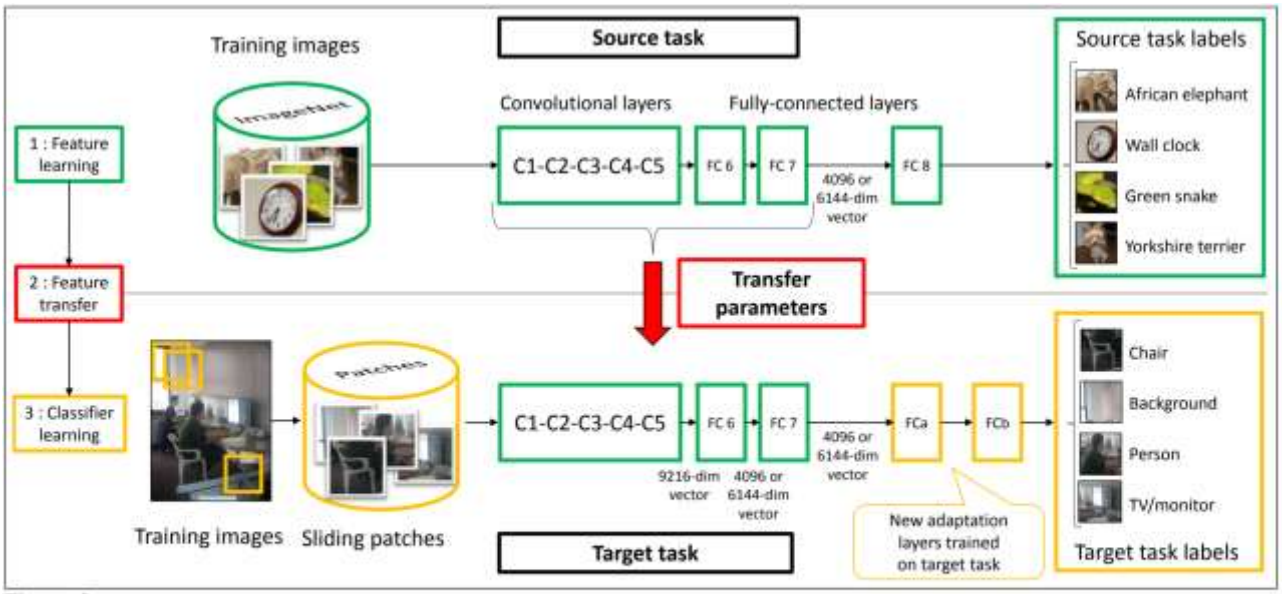


Figure 2: **Transferring parameters of a CNN.** First, the network is trained on the source task (ImageNet classification, top row) with a large amount of available labelled images. Pre-trained parameters of the internal layers of the network (C1-FC7) are then transferred to the target tasks (Pascal VOC object or action classification, bottom row). To compensate for the different image statistics (type of objects, typical viewpoints, imaging conditions) of the source and target data we add an adaptation layer (fully connected layers FCa and FCb) and train them on the labelled data of the target task.

L'apprentissage par transfert (transfer learning) permet aussi de pré-entraîner un réseau de reconnaissance d'images avec des images générées synthétiquement dans diverses positions[170]. On l'utilise aussi pour améliorer la labellisation d'illustrations[171].

Descente stochastique de gradient (2003)

La SGD (Stochastic Gradient Descent) est une technique d'apprentissage par rétropropagation des erreurs qui s'appuie sur l'optimisation du gradient. Pour faire simple, il s'agit d'identifier dans quelle direction faire évoluer les poids synaptiques des neurones pour atteindre leur niveau optimal dans la reconnaissance des objets en minimisant les opérations de calcul nécessaires. Le tout étant utilisé dans l'entraînement du réseau de neurones par rétropropagation d'erreurs.

Il est important de trouver le niveau optimum global, à savoir le taux d'erreur le plus bas, et pas seulement le niveau optimum local, qui est le taux d'erreur le plus bas dans les environs du poids de départ que le réseau de neurones cherche à optimiser (voir le schéma *ci-dessous* qui l'explique de manière imagée).

La technique s'applique aussi bien aux réseaux de neurones à une seule couche cachée qui font partie du domaine du machine learning qu'aux réseaux de neurones complexes du deep learning.

---

[169] Source de l'illustration : Learning and Transferring Mid Level Image Representations using Convolutional Neural Networks par Maxime Oquab (Inria) et al, 2014 (8 pages).

[170] Voir par exemple Improving SAR Automatic Target Recognition Models with Transfer Learning from Simulated Data de David Malmgren-Hansen, Anders Kusk, Jørgen Dall, Allan Aasbjerg Nielsen, Rasmus Engholm et Henning Skriver, 2017 (6 pages).

[171] Voir Transfer Learning for Illustration Classification, 2018 (9 pages).

Dans l'entraînement par rétropropagation d'erreurs, les poids synaptiques des neurones sont initialisés aléatoirement. On fait passer des objets d'une base de test au travers du réseau et on compare le résultat de classification en sortie avec le bon résultat dont on dispose dans la base d'entraînement (en amont, des photos et en aval, des descripteurs des objets dans les photos).

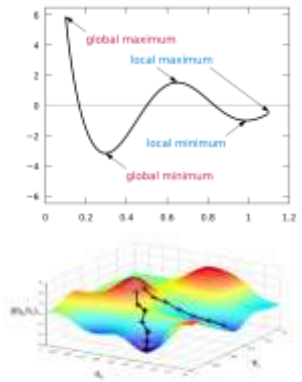
La descente de gradient évalue dans quelle direction faire évoluer les poids des synapses pour s'approcher du bon résultat. Le calcul est réalisé pour toutes les synapses et pour tous les objets du jeu d'entraînement, ce qui génère beaucoup de calculs[172].

La descente stochastique de gradient est une variante de la descente de gradient qui consiste à évaluer le poids des synapses objet par objet (méthode stochastique) ou groupe d'objets (méthode mini-batch) du groupe d'objets d'entraînement au lieu de scanner entièrement la base d'entraînement (méthode dite « batch »). Cela permet de réduire la quantité de calculs à effectuer et permet de converger plus rapidement vers un réseau bien entraîné.

Cette technique d'entraînement est très efficace pour générer un réseau de neurones capable de générer des résultats avec un faible taux d'erreurs.

Elle est cependant très consommatrice de ressources machines et de temps.

D'où l'intérêt de l'optimiser et de s'appuyer sur des ressources matérielles de plus en plus puissantes, comme les ordinateurs à base de GPU ou de processeurs neuromorphiques que nous étudions plus loin dans ce document.



faire varier les poids des neurones pour trouver la valeur optimale globale, celle qui minimise le niveau d'erreur du réseau de neurones.

opération effectuée pour chaque neurone et pour chaque objet de la base de référence.

"stochastic gradient descent" est une des optimisations qui ajuste les paramètres du réseau de neurones objet par objet ou par groupe d'objet au lieu de les calculer d'un coup pour tous les objets d'entraînement à la fois

la difficulté consiste à trouver le minimum global et pas simplement un minimum local pour chaque descente de gradient de chaque poids de synapse pour chaque objet de la base d'entraînement

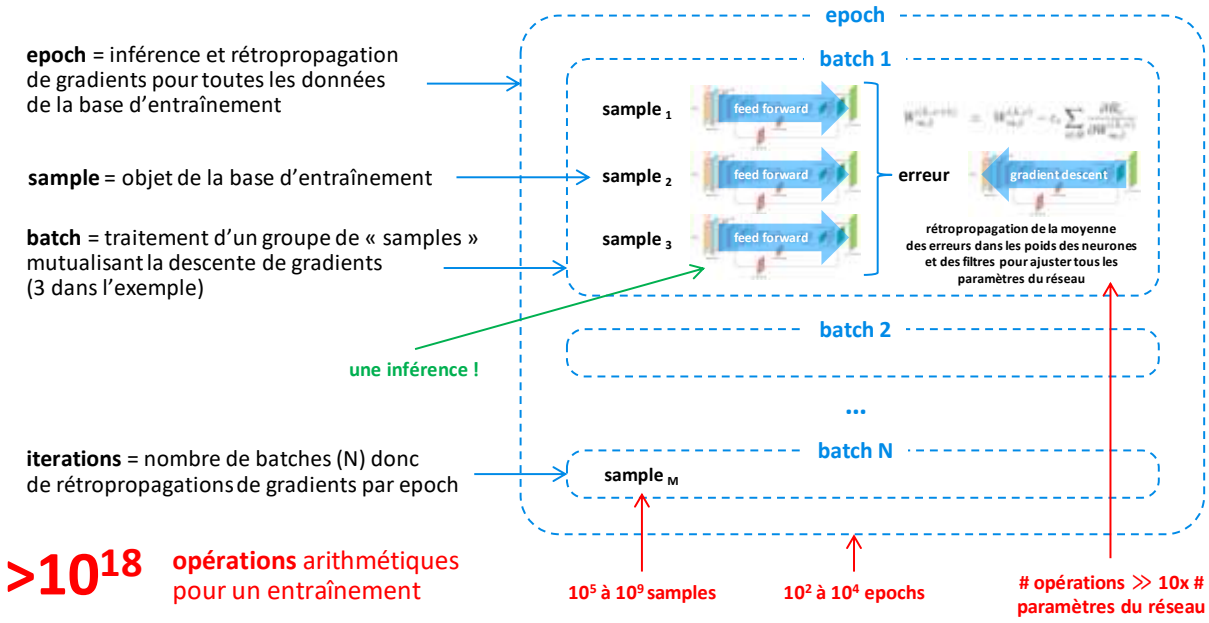
Voici quelques éléments de dimensionnement et de jargon de l'entraînement d'un réseau de neurones lié à l'application d'une SGD[173] :

- Un **sample** est un objet de la base d'entraînement.
- Ils sont regroupés dans des **batches**, la rétropropagation des erreurs étant réalisée sur la moyenne des erreurs générées par une forward pass dans le réseau de neurones avec les samples les composant.
- Le passage de tous les samples à la moulinette précédente se fait en un nombre d'**itérations** représentant le nombre de batches.
- Une **epoch** correspond à un passage de l'ensemble des échantillons dans le circuit. À savoir une *forward pass* d'inférence et autant de *backward passes* de rétropropagation des gradients que de *batches*.
- On itère plusieurs fois le processus avec un nombre d'epoch qui se compte en centaines voire en milliers.

[172] Voir Lecture 7: Training Neural Networks, Part 2, 2018 (110 slides) qui décrit dans le détail les méthodes d'entraînement des réseaux de neurones, notamment l'initialisation des poids des neurones et la descente stochastique de gradient. Voir aussi Neural Networks and Introduction to Deep Learning (17 pages) qui décrit bien les modèles mathématiques de l'entraînement de réseaux de neurones et en particulier la descente stochastique de gradient.

[173] Voir Difference Between a Batch and an Epoch in a Neural Network par Jason Brownlee, juillet 2018.

À la fin et pour prendre l'exemple d'un entraînement d'un réseau ResNet-50 sur 90 epochs, il faut exécuter la bagatelle de 10^{18} opérations. Cela peut prendre de quelques jours sur un PC équipé d'un bon GPU Nvidia équipé d'unités de calcul de tenseurs jusqu'à quelques minutes sur les plus grands supercalculateurs, en exploitant des méthodes de parallélisation de l'entraînement[174]. En pratique, le nombre d'opérations à réaliser pour un entraînement est de plusieurs ordres de grandeur plus élevé.

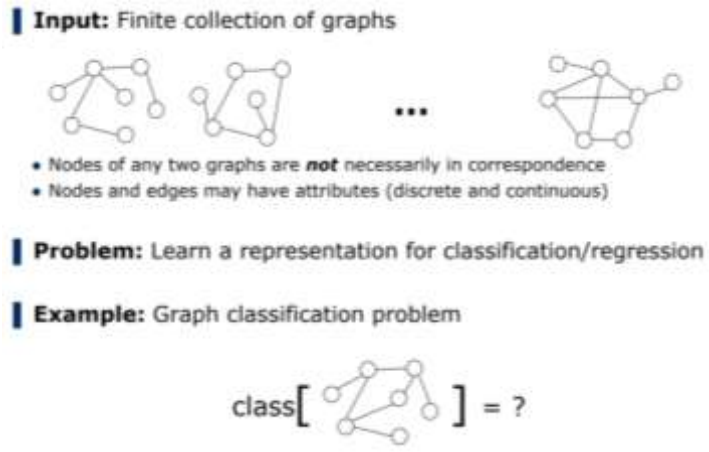


Graph Neural Networks (2005)

Les CNN et RNN sont des réseaux de neurones profonds qui analysent des données structurées hiérarchiquement dans ce que l'on appelle un espace euclidien généralement à une ou deux dimensions (1D ou 2D). Les graphes comprennent des objets qui sont reliés entre eux dans une organisation arbitraire et non hiérarchique. Ils sont structurés avec des éléments (points) et de relations entre ces points (qui peuvent être des labels). Ils peuvent être organisés en 2D ou 3D comme sur une sphère ou un tore.

Les Graph Neural Networks servent servent à résoudre des problèmes liés à des structures de données non spatiales au sens « pixels » du terme[175] (images, sons, bruits).

On les exploite notamment pour gérer des systèmes industriels complexes, des bases de molécules organiques et leurs caractéristiques, les effets de la combinaison de médicaments ainsi que pour de l'analyse de code dans les outils de développement, etc[176].

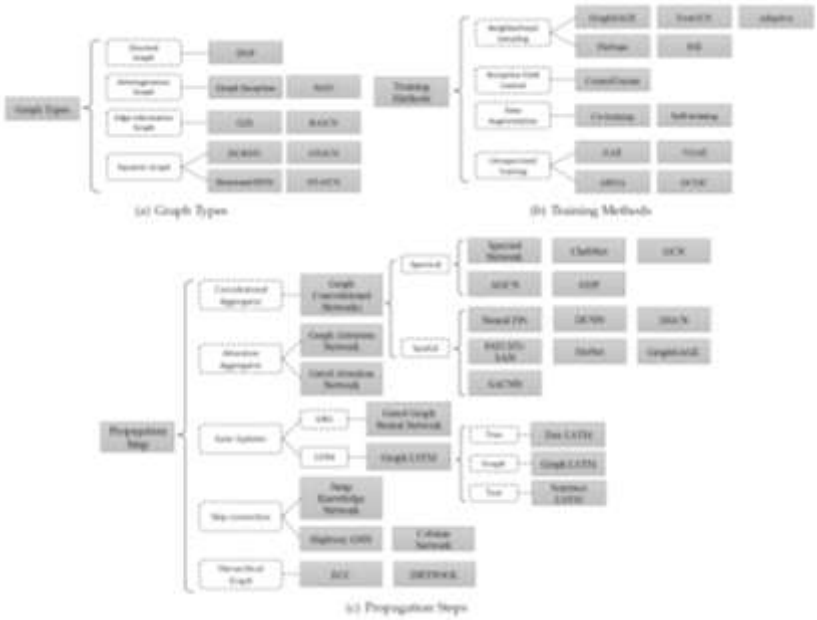


[174] Voir ImageNet Training in Minutes par Yang You et al, 2017 (11 pages).

[175] Voir Convolutional Neural Networks on Graphs par Xavier Bresson, février 2017 (87 slides) et la vidéo associée qui décrit très bien les différences conceptuelles entre les réseaux de neurones convolutifs et les réseaux de neurones de graphes. Puis Convolutional Neural Networks on Graphs par Xavier Bresson, juin 2017 (71 slides), Graph Neural Networks (66 slides), Graph Neural Networks par Alejandro Ribeiro, juin 2019 (64 slides), Structured deep models: Deep learning on graphs and beyond par Thomas Kipf, mai 2018 (100 slides), Deep Learning on Graphs with Graph Convolutional Networks par Thomas Kipf, mars 2017 (52 slides), Simplifying Graph Convolutional Networks (11 pages) et How Powerful Are Graph Neural Networks ? par Keyulu Xu & Al, 2019 (17 pages).

Ils ont été conçus initialement entre 2005 et 2009[177]. Il en existe de nombreux types types qui continuent encore d'évoluer aujourd'hui (schéma *ci-dessous*) [178].

Dans le principe, un graphe relie des nœuds entre eux via des liens quantifiés ou labellisés. Le GNN va permettre d'entraîner une fonction f(graphe) à générer un label pour l'ensemble du graphe ou pour des nœuds ou liens de graphes. On entraîne le GNN avec plusieurs graphes d'une base (comme des molécules avec les atomes reliés entre eux dans le graphe). La base d'entraînement est une série de graphes avec leur valeur f(graphe). Le réseau de neurones entraîné va permettre d'évaluer f(graphe) pour un graphe sans son label/sa valeur.



Certaines méthodes de classification des graphes consistent à transformer ces graphes en matrices 2D ou 3D, les cases de ces matrices représentant les liens entre deux nœuds du graphe et leur type de lien (ce qui fait trois dimensions) et à appliquer dessus un ConvNet[179].

D'autres types de GNN ont comme objet de labelliser un nœud ou un lien du graphe dont on cherche le label.

Enfin, il existe aussi des GNN génératifs (de type GAN : generative adversarial networks) pouvant servir à créer de nouveaux graphes à partir de graphes existants, comme pour créer de nouvelles molécules thérapeutiques.

Les domaines d'applications des GNN sont très variés et sont en pleine croissance. En premier lieu, on trouve la chimie pour prédire les fonctions de molécules, pour créer des systèmes de reconnaissance de nouvelles classes d'objets dans les images, ainsi que pour détecter des structures grammaticales dans le traitement du langage[180].

Deep beliefs networks (2006)

Les DBN sont issus des travaux des canadiens Geoffrey Hinton et Simon Osindero et du Singapourien Yee-Whye Teh[181]. Ils optimisent le fonctionnement des réseaux neuronaux multicouches en gérant leur apprentissage couche par couche, indépendamment les unes des autres.

---

[176] Source de l'illustration : Learning Convolutional Neural Networks for Graphs par Mathias Niepert (38 slides).

[177] Voir A new model for learning in graph domains, M. Gori, G. Monfardini, and F. Scarselli, 2005 (7 pages) et The graph neural network model par Franco Scarselli & Al, 2009 (22 pages).

[178] Voir Graph Neural Networks: A Review of Methods and Applications par Jie Zhou & Al, juillet 2019 (22 pages) et A Comprehensive Survey on Graph Neural Networks par Zonghan Wu & Al, août 2019 (22 pages).

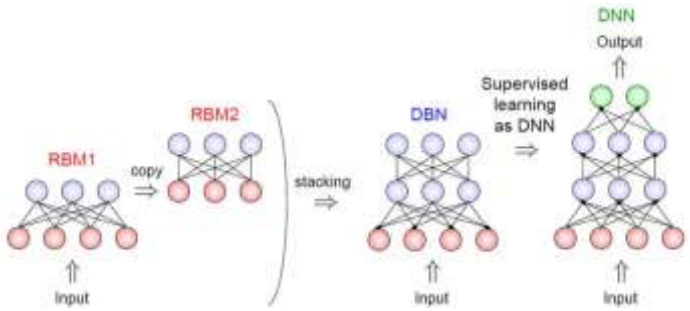
[179] Voir la présentation Learning the Structure of Graph Neural Networks par Mathias Niepert, juillet 2019 (1h27mn) ainsi que Steve Purves - Graph Convolutional Networks for Node Classification, mars 2018 (32 minutes).

[180] Voir Applications of Graph Neural Networks par Aishwarya Jadhav, février 2019.

[181] Voir A Fast Learning Algorithm For Deep Belief Nets, 2006 (16 pages).

Ce sont en quelque sorte des machines de Boltzmann restreintes empilées les unes sur les autres, étape par étape pour ce qui est de l'entraînement.

Le concept général du deep learning a été ensuite formalisé par ce même Geoffrey Hinton en 2007[182].



Notons que Geoff Hinton s'appuyait en partie sur les travaux du français **Yann Le Cun** (en 1989) qui dirige maintenant le laboratoire de recherche en IA de Facebook et de l'allemand **Jürgen Schmidhuber** (1992, créateur des LSTM) dont deux des anciens étudiants ont créé la start-up **DeepMind**, maintenant filiale de Google. Sachant que Yann Le Cun était lui-même un ancien post-doctorant dans le laboratoire de Geoff Hinton. Un bien petit monde !

Geoffrey Hinton[183] travaille pour **Google** depuis 2013, pas loin du légendaire **Jeff Dean**[184], arrivé en 1999 et qui planche maintenant aussi sur le deep learning**.** On peut aussi citer le français **Stéphane Mallat** qui a aussi contribué au développement des réseaux convolutifs et à l'explication de leurs sous-jacents mathématiques[185].

Knowledge Distillation (2006)

Il s'agit de méthodes d'entraînement de réseaux de neurones à partir de réseau de neurones existants, dont ils sont des sortes de sous-ensembles, à différencier de l'apprentissage par transfert. Cela permet d'être plus efficace que d'entraîner directement le réseau de neurone plus simple.

Autoencodeurs empilés (2007)

Les *stacked autoencoders* sont couramment associés aux deep belief networks et aux réseaux convolutifs. Il s'agit d'utiliser des couches cachées de neurones qui encodent les données dans un espace dimentionnel réduit et des couches de neurones qui permettent ensuite de reconstituer les variables en entrées, en sortie de cette couche.

Cette technique est utilisée dans l'apprentissage non supervisé des réseaux de neurones pour identifier des variables ou fonctions cachées. Elle peut notamment servir à débruiter des images.

Cette technique est donc une technique de réduction de dimensions. La méthode mathématique employée peut être la PCA (Principal Components Analysis) que nous avons rapidement vue dans la partie sur le machine learning ou la Singular Value Decomposition (SVD), autre méthode voisine de la PCA. En fait l'autoencodeur réalise une sorte de PCA non linéaire[186].

Il existe bien entendu diverses variantes d'autoencodeurs comme les DAE (denoising autoencoders), les SAE (sparse autoencoders) et les VAE (variational autoencoders).

---

[182] Voir Learning multiple layers of representation, 2007 (100 slides).

[183] Voir Is AI Riding a One-Trick Pony? de James Somers, septembre 2017, MIT Technology Review, qui montre à quel point Geoff Hinton est central dans l'histoire récente de l'IA.

[184] Jeff Dean est le co-créateur, entre autres choses, de deux outils clés des traitements distribués MapReduce et BigTable, ainsi que du crawler de Google Search et d'AdSense.

[185] Sa conférence délivrée dans la Chaire du Collège de France de Yann Le Cun fournit des éclaircissements sur le fonctionnement des réseaux convolutifs. Mais il faut s'accrocher pour suivre ! Voir les Mystères mathématiques des réseaux de neurones convolutionnels, 19 février 2016.

[186] Voir Auto-association by multilayer perceptrons and singular value decomposition, de Hervé Bourlard et Y. Kamp, 1988.

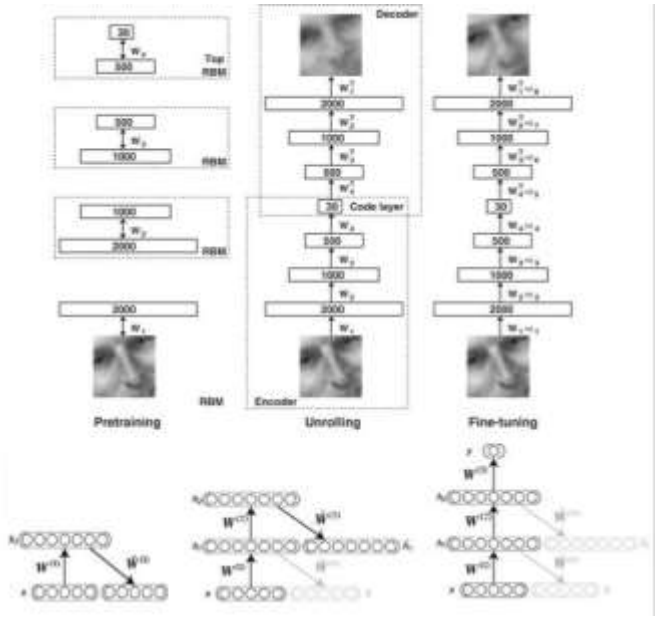
les autoencodeurs empilés sont des couches cachées de neurones qui encodent les données dans un espace dimentionnel réduit

l'autoencodeur est capable de reconstituer les variables en entrées en sortie de cette couche

utilisé en mode d'apprentissage non supervisé pour trouver des variables cachées, y compris dans les CNN, la couche de réencodage n'étant pas utilisée en production

peut notamment servir à débruiter des images (de caractères, ...) dans des autoencodeurs empilés

la méthode employée peut-être la PCA (Principal Components Analysis)

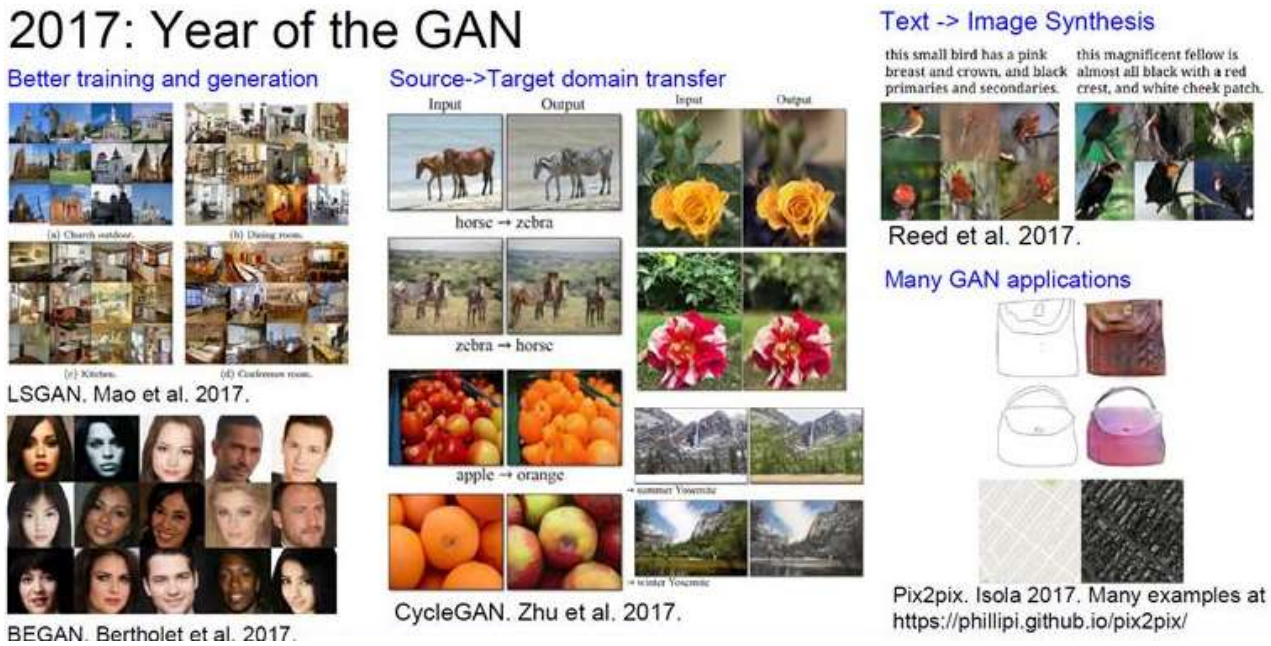


Réseaux de neurones génératifs *(2014)*

Les réseaux de neurones génératifs sont des réseaux de neurones convolutifs ou récurrents générant du contenu à partir de contenus existants.

Ils font des prévisions, d'images vidéo suivantes d'une vidéo donnée, ils colorient des images en noir et blanc[187] et ils remplissent des images dont il manque des morceaux. Ils peuvent aussi servir à améliorer les dialogues de chatbots.

Les principales techniques utilisées sont les **GAN** (Generative Adversarial Networks), apparus en 2014 et perfectionnés en particulier en 2016. Ces sont des réseaux de neurones non supervisés capables de générer des contenus en modifiant une ou plusieurs variables latentes dans des générateurs à base de réseaux de convolution inversés, sortes de stacked autoencoders multiples. Un discriminateur utilisant un réseau de neurones dit critique permet alors d'identifier les contenus générés les plus plausibles, en vérifiant que les objets générés sont bien plausibles.



[187] Les exemples du slide ci-dessous viennent de : Generative Models de Fei-Fei Li & Justin Johnson & Serena Yeung, 2017.

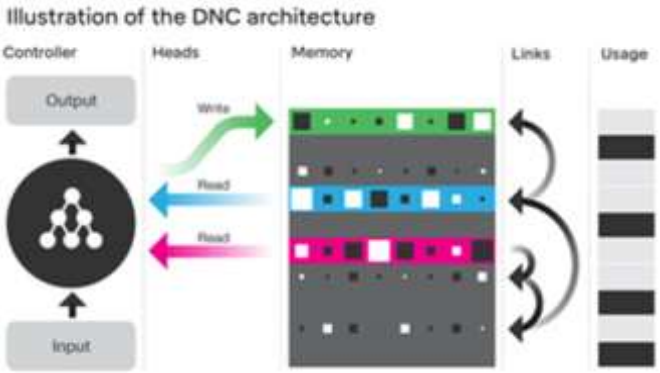
Les variantes de réseaux génératifs sont de plus en plus nombreuses avec notamment les **VAE** (Variational Autoencoders, 2014), les **Adversarial Eutoencoders** (2015), les **InfoGAN** (2016), les **CycleGAN** (2017[188]), **Wasserstein GAN** (2017), **PixelRNN / PixelCNN** (2016), **GMMN** (Generative moment matching networks, 2015) et des versions évolutionnaires comme **E-GAN** (2018)[189].

Je traite des GANs dans la rubrique sur les modèles génératifs dans la partie liée à la vision car c'est le principal domaine d'application des GANs[190]. Nous les verrons aussi dans la rubrique spécifique dédiée aux fake news.

Differential Neural Computers (2016)

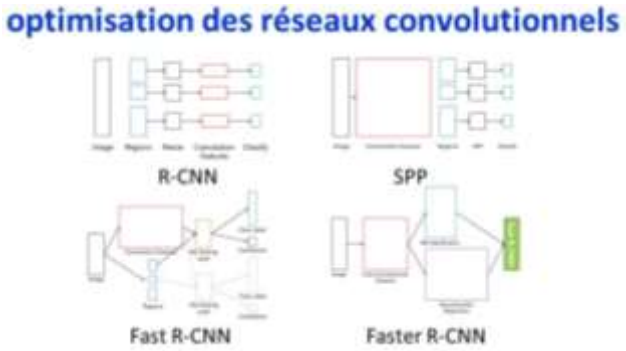
Les DNC sont des réseaux de neurones récurrents utilisant une mémoire autoassociative, créés par Alex Graves et son équipe de DeepMind[191]. Ils gèrent les relations entre composantes d'une mémoire long terme.

A ce jour, ils ne servent pas à grand-chose mais pourraient servir à intégrer du raisonnement symbolique dans les réseaux de neurones. Les DNC peuvent répondre à des questions sur des données structurées complexes et structurées en arbres et même résoudre certains types de puzzles. On associe ces DNC à la programmation différentielle dont l'usage est actuellement en plein essor[192]. Ils imitent aussi le fonctionnement de l'hippocampe du cerveau humain.



Autres méthodes de deep learning (2015-2020)

S'en suivirent plus récemment de nombreuses variantes de réseaux de neurones, surtout à base de réseaux convolutifs (*ci-dessous à droite*), destinées à optimiser les performances, en particulier, celles de l'entraînement des réseaux. En effet, c'est la partie qui est la plus consommatrice de ressources machines dans les réseaux de neurones. Une fois un réseau entraîné, il exécute ses prévisions bien plus rapidement.



On voit aussi émerger des réseaux de **deep learning évolutifs** dont l'architecture peut évoluer de manière itérative[193].

Le schéma *ci-dessous* illustre cette longue chaîne de progrès, qui ne s'est d'ailleurs pas arrêtée en 2012 et poursuit encore son chemin aujourd'hui.

---

[188] Voir Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks par Jun-Yan Zhu, 2017 (18 pages). L'une des démonstrations consiste à transformer un cheval en zèbre et réciproquement.

[189] Evolutionary Generative Adversarial Networks par Chaoyue Wang & Al, 2018 (14 pages).

[190] Voir aussi An applied introduction to generative adversarial networks, décembre 2017.

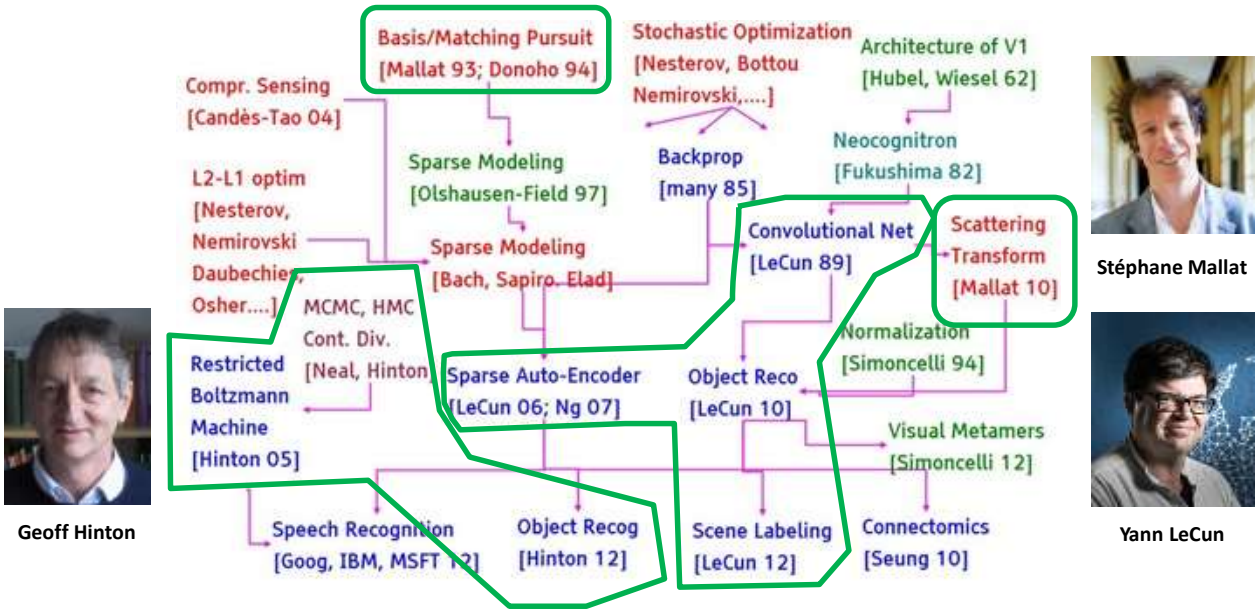
[191] Voir Hybrid computing using a neural network with dynamic external memory, 2016 (23 pages), Differentiable Neural Computers d'Alex Graves, 2016 (16 slides) et Differentiable neural computers, 2016, du même auteur, et What are differentiable neural computers de Heidelberg.ai, 2017 (62 slides). Sinon, Hardware Accelerator for Differentiable Neural Computer and Its FPGA Implementation, 2017 (7 pages) décrit la manière de transposer les DNC dans du hardware.

[192] Voir notamment Diffential programming de Atilim Güneş Baydin, 2016 (69 slides).

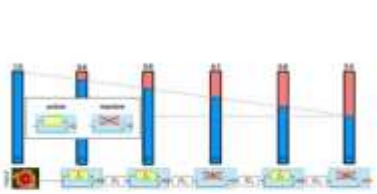
[193] Voir Neuroevolution: A different kind of deep learning, juillet 2017.

Elle s'est même accélérée avec l'arrivée dans le domaine de très nombreux chercheurs chinois. 2012 est une date intéressante, celle de la création du réseau de neurones convolutif **AlexNet** par l'équipe de Geoff Hinton de l'Université de Toronto. Il était entraîné sur 1,3 millions d'images de la base de référence **ImageNet** avec 1000 classes d'images différentes et générait des taux d'erreurs bas pour l'époque (18,9% pour le top 5).

Le réseau de neurones exploitait 60 millions de paramètres et 500 000 neurones répartis dans cinq couches de convolution. Le tout exploitait des cartes GPU Nvidia pour accélérer les traitements, surtout pendant la phase d'entraînement qui durait des semaines[194].

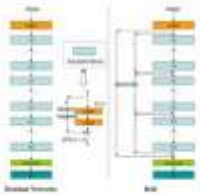


Le deep learning et les réseaux de neurones convolutifs ont continué de progresser pas à pas, avec la création d'innombrables variantes conceptuelles pour améliorer leurs capacités d'apprentissage, de mémorisation et leur performance d'entraînement. L'IA progresse d'ailleurs régulièrement et de manière plutôt décentralisée, avec des dizaines de chercheurs contribuant à faire avancer l'état de l'art. Le tout, pendant que les GPU et processeurs neuromorphiques progressent tout autant de leur côté.

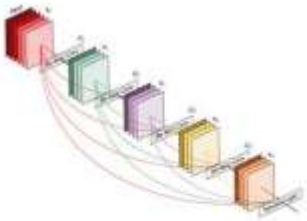


Stochastic Residual Net, 2016
accélère entrainement

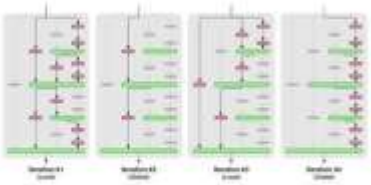
Residual Net, 2016
duplication de couches,
améliore entrainement



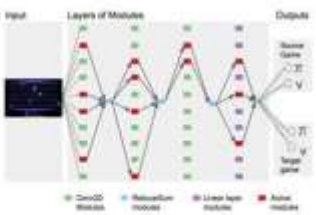
RoR, 2016
variante de ResNet



DenseNet, 2016
évite perte de gradient



FractalNel, 2016
densification de réseau

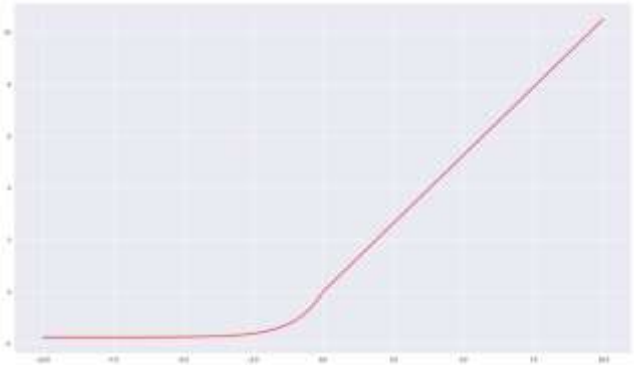


PathNet, 2016
réseau de réseau

[194] Voir ImageNet Classification with Deep Convolutional Neural Networks, 2012.

Nous avons ainsi par exemple vu apparaître quelques avancées conceptuelles clés depuis 2015 :

- **ResNet**[195] (2015), un réseau de neurones profonds pouvant avoir jusqu'à 152 couches, mais qui en optimise l'entraînement via des court-circuitage de couches de convolution. Il sert à réduire la perte de gradients dans le processus d'entraînement. Il permettait d'atteindre un taux d'erreur de 5,7% sur la base de tests ImageNet. La variant ResNet-50 à 50 couches dont 48 de convolution est la plus couramment utilisée.
- **Stockastic Residual Net**[196] (2016), qui optimise les réseaux de neurones en court-circuitant certaines couches pendant l'entraînement pour le rendre plus rapide.
- **FractalNet**[197] (2016), qui utilise le concept des fractales pour densifier un réseau de neurones convolutionnel en répliquant certaines couches et en utilisant plusieurs circuits différents pour l'optimisation de chaque convolution.
- **DenseNet**[198] (2016), une variante des ConvNets où chaque feature map est injectée en entrée de toutes les couches de convolution suivantes et pas seulement de la suivante, évitant le syndrome de la perte de gradient qui affecte les ConvNets lors de leur entraînement.
- **PathNet**[199] (2016), un réseau de neurones, chaque neurone étant un réseau convolutif, dont l'usage est optimisé automatiquement. C'est une création de DeepMind.
- **SqueezeNet** (2016) est une version compacte de réseaux de neurons convolutifs qui est très économe en mémoire et exploitable dans les systèmes embarqués qui sont contraints en ressources[200].
- **SNN** (2017) sont les Self-Normalizing Neural Networks[201], des variantes de ConvNets qui utilisent une function d'activation particulière, le **SELU** (Scaled Exponential Linear Units) qui intègre une fonction de normalisation qui n'est pas très éloignée d'une RELU et d'une ELU en étant légèrement négative pour les valeurs négatives au lieu d'être à zéro[202]. Qui plus est, les valeurs des poids sont initialisées à 0 a lieu de 0,5. Pourquoi donc ? Pour permettre une convergence plus rapide du réseau de neurones lors de l'entraînement par rétropropagation.



Une fonction d'activation SELU permet d'améliorer d'un facteur 10 le temps d'entraînement de certains types de réseaux de neurones convolutifs.

---

[195] Voir Deep Residual Learning for Image Recognition, de Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun, 2015. ResNet a été développé par une équipe de Microsoft Research en Chine. ResNet serait utilisé dans AlphaGo Zero, développé par DeepMind en 2017.

[196] Voir Deep Networks with Stochastic Depth de Gao Huang, Yu Sun, Zhuang Liuy, Daniel Sedra, et Kilian Q. Weinberger, 2016.

[197] Voir FractalNet : Ultra-deep neural networks without residuals, de Gustav Larsson, Michael Maire et Gregory Shakhnarovitch, 2016.

[198] Voir Densely Connected Convolutional Networks de Gao Huang, Zhuang Liu et Laurens van der Maaten, 2016, révisé en 2017.

[199] Voir PathNet: Evolution Channels Gradient Descent in Super Neural Networks, de Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel et Daan Wierstra, janvier 2017.

[200] Voir SqueezeNet: AlexNet-Level accuracy with 50X fewer parameters and <0.5MB model size, par Forrest N. Iandola & Al de le société DeepScale (qui créé des logiciels pour les systèmes de conduite assistée et autonome et des Universités de Stanford et Berkeley, novembre 2016 (13 pages). SqueezeNet remplace notamment des filtres 3x3 par des filtres 1x1.

[201] Voir Self-Normalizing Neural Networks de Günter Klambauer, Thomas Unterthiner et Andreas Mayr (Autriche), 2017 (102 pages).

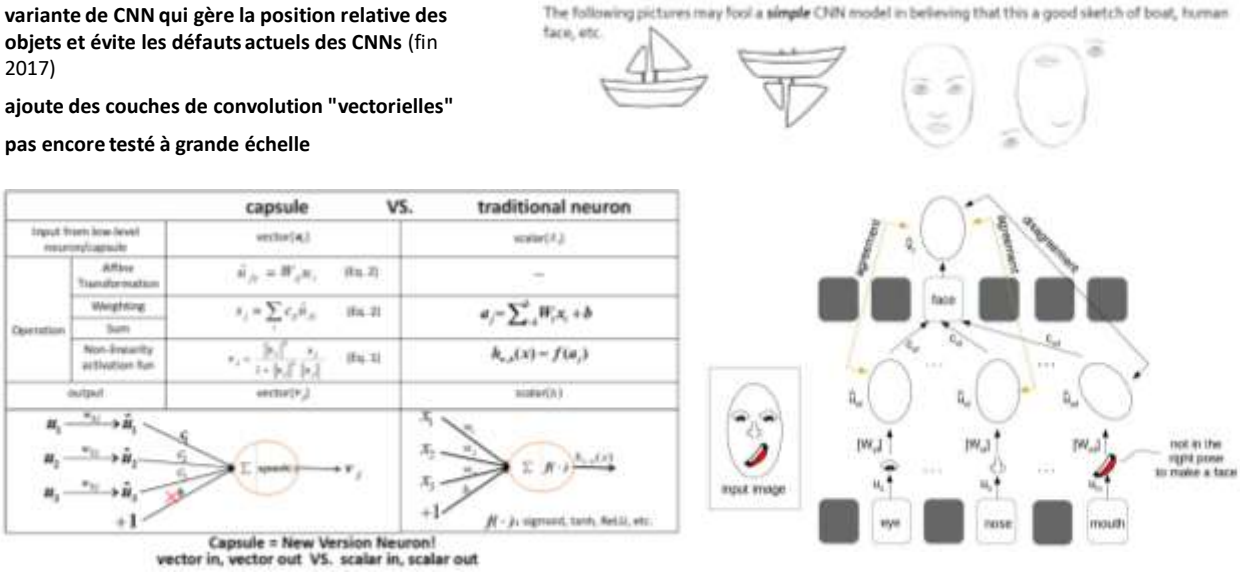
[202] Voir Normalization Techniques in Deep Neural Networks par Aakash Bindal, février 2019. Qui explique les techniques de normalisation dans les réseaux de neurones.

- **Capsule Networks** (2017) ont été présentés par Geoff Hinton[203] comme un moyen de permettre aux réseaux convolutifs de ne pas perdre la spatialité relative des objets intermédiaires qu'ils découvrent. Ils complètent les convnets avec des couches vectorielles gérant la position relative des sous-objets découverts dans chaque couche de convolution. Pour l'instant, les Capsule Networks n'ont été testés que pour reconnaître des lettres et ils présentent l'inconvénient de nécessiter un très grand nombre de paramètres rendant leur entraînement peut-être difficile à réaliser sur les processeurs et mémoires d'aujourd'hui. Des déclinaisons des Capsule Networks ont depuis été proposées pour améliorer la classification d'objets[204], celle de tumeurs cancéreuses du cerveau[205] et pour l'analyse de sentiments[206]. Il ne semble pas que ces réseaux percent véritablement.

variante de CNN qui gère la position relative des objets et évite les défauts actuels des CNNs (fin 2017)

ajoute des couches de convolution "vectorielles"

pas encore testé à grande échelle



- **Mixture of Expert Layer**[207] (2017) est un modèle de réseau de neurones multicouches créé par une équipe de Google Brain pilotée par Geoff Hinton. C'est un réseau neuronal géant dont chaque neurone est en fait un sous-réseau neuronal. Le modèle, différent de Pathnet, sert surtout à améliorer les outils de traitement du langage comme la traduction.
- **Tree-CNN** (2018) est une des méthodes les plus récentes permettant l'entraînement incrémental d'un Convnet[208]. En effet, comme un Convnet demande d'énormes ressources machines pour être entraîné, des méthodes spécifiques sont nécessaires pour modifier cet entraînement lorsque, par exemple, des images sont ajoutées voire même supprimées d'un jeu de test.
- **MobileNet** (2018) est une autre forme de réseaux convolutifs légers ciblant les systèmes embarqués, créé par Google[209].

---

[203] Voir Dynamic Routing Between Capsules de Sara Sabour, Nicholas Frosst et Geoffrey Hinton, novembre 2017 (11 pages), cette vidéo d'Aurélien Géron qui en décrit bien le principe, 'Godfather' of deep learning is reimagining AI de Chris Sorensen, novembre 2017 et cette vulgarisation technique de haut vol Understanding Hinton's Capsule Networks en quatre partie de Max Pechyonkin.

[204] Voir Capsules for Object Segmentation de Rodney LaLonde et Ulas Bagci, avril 2018 (9 pages).

[205] Voir Brain tumor type classification via capsule networks de Parnian Afshar, Arash Mohammadi et Konstantinos Plataniotis, mars 2018 (5 pages).

[206] Voir Sentiment Analysis by Capsules, avril 2018 (10 pages).

[207] Voir Outrageously large neural networks : the sparsely-gated mixture-of-experts layer de Geoffrey Hinton, Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le et Jeff Dean, janvier 2017.

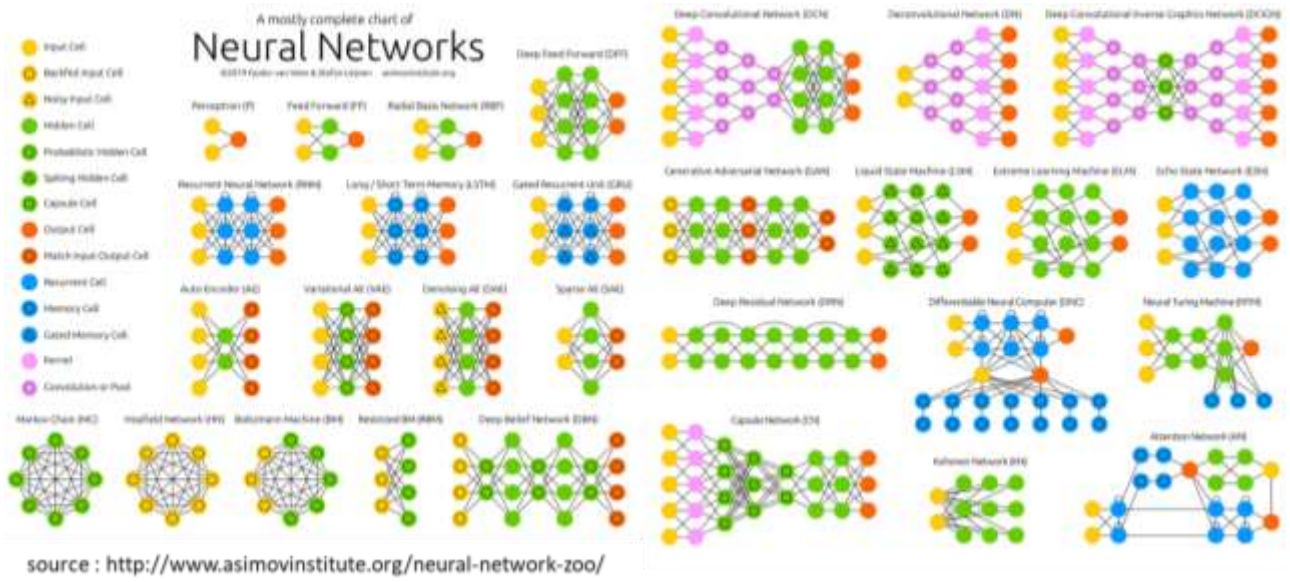
[208] Voir A Hierarchical Deep Convolutional Neural Network for Incremental Learning, mai 2018 (12 pages).

[209] Voir MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications par Andrew G. Howard, avril 2017 (9 pages).

- **Gauge Equivariant CNNs** (2019) associe réseaux de neurones convolutifs et topologie pour permettre leur application sur des surfaces complexes comme des sphères comme une mappemonde ou des formes topologiques diverses et variées comme des icosahédrons[210].
- **EfficientNet** (2019), une variante de réseau convolutif plus dense qui améliore la précision et l'efficacité des inférences[211].
- **ODE Net** (2019) pour Ordinary Differential Equations Net est une technique de réseaux de neurones sans couches de convolution qui permet de détecter des features avec plusieurs niveaux d'abstraction[212]. Il remplace les couches par un vortex de neurones reliés entre eux par des calculs continus. C'est très abstrait et l'histoire n'indique pas à quoi cela peut servir.
- **DeepSpeed** (2020) est une bibliothèque open source de deep learning écrite en PyTorch lancée par Microsoft Research. Elle permet la distribution de l'entraînement d'un réseau de neurones sur plusieurs processeurs, surtout des GPU. Il peut dépasser la centaine de milliards de paramètres. Ce projet fait partie de l'initiative « AI at scale » de Microsoft et s'applique en particulier aux modèles de traitement du langage.

À chaque fois, ces différents réseaux ont été entraînés avec les mêmes sources de données comme la base **ImageNet**, pour reconnaître avec le taux d'erreurs le plus faible les images de test et aussi en économisant au mieux les ressources machine. L'autre point commun de ces avancées récentes est qu'elles proviennent souvent de chercheurs et étudiants chinois… installés surtout aux USA.

Depuis 2016, les taux d'erreurs sont inférieurs à ceux de l'Homme ce qui explique pourquoi il est souvent dit qu'en matière d'imagerie médicale, les médecins spécialistes sont dépassés. Même si ce propos est exagéré car la reconnaissance d'images ne correspond qu'à une partie seulement de leur expertise. A ceci près que ces réseaux de neurones ont été entraînés avec des bases d'images taggées reflétant le savoir existant des spécialistes. La connaissance de l'IA ne tombe pas du ciel !



source : http://www.asimovinstitute.org/neural-network-zoo/

[210] Voir An Easy Guide to Gauge Equivariant Convolutional Networks par Michael Kissner, mai 2019 et l'article d'origine Gauge Equivariant Convolutional Networks and the Icosahedral CNN par Taco S. Cohen et. Al, février 2019 (15 pages).

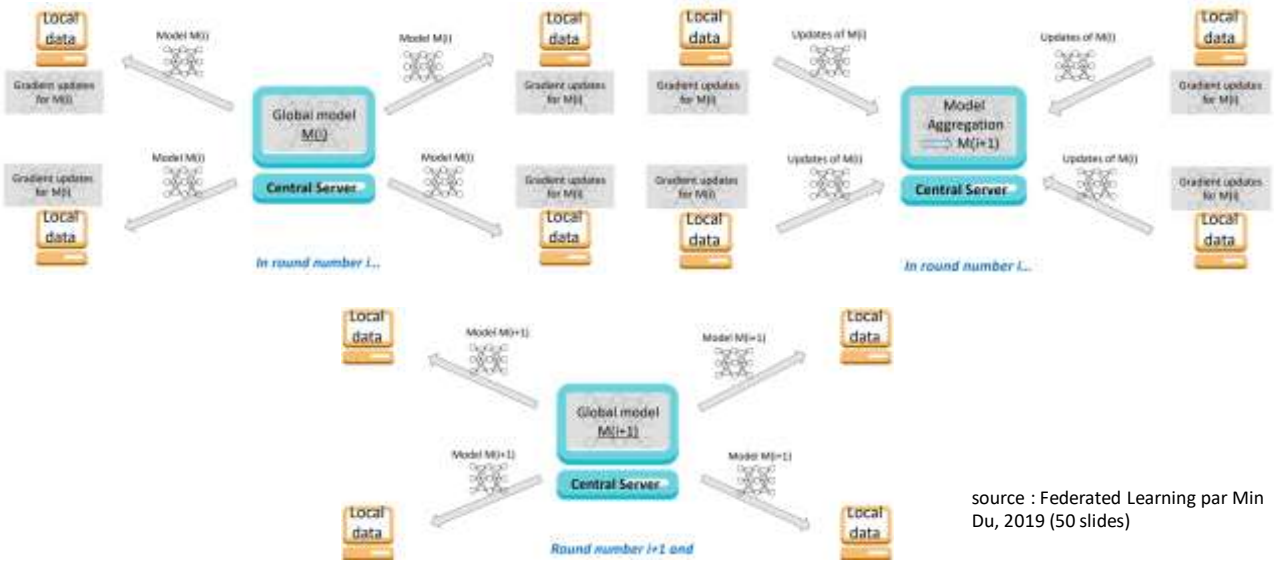
[211] Voir EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks par Mingxing Tan et Quoc V. Le, 2020 (11 pages).

[212] Voir A radical new neural network design could overcome big challenges in AI par Karen Hao, décembre 2018 qui fait référence à Neural Ordinary Differential Equations par David Duvenaud et al, juin 2018 (19 pages).

La cartographie *ci-dessus* du « zoo » des réseaux de neurones[213] illustre bien leur diversité sachant que leur assemblage peut donner ensuite lieu à beaucoup de créativité en fonction des besoins.

Federated learning *(2017-2019)*

Le federated learning est un modèle d'entraînement distribué de réseaux de neurones de deep learning applicable notamment à la reconnaissance d'image et au traitement du langage. Il sert plusieurs objectifs, notamment en relation avec les usages mobile. Le principal est de protéger la vie privée des utilisateurs et des données qu'ils génèrent qui peuvent servir à entraîner des modèles de deep learning de manière incrémentale. Le second est aussi de mieux répartir les traitements. Vu de haut, le principe est simple. Il consiste à réaliser une partie de l'entraînement en local comme sur un smartphone puis d'envoyer des éléments de modèle entraînés au datacenter qui va consolider l'ensemble mais sans avoir eu accès aux données d'entraînement. Le datacenter va ensuite distribuer le modèle amélioré de manière incrémentale à tous les utilisateurs. Ce genre de modèle est par exemple utilisé par Google pour l'auto-remplissage de textes saisis dans les applications mobiles Android. Un autre de ses bénéfices est de réduire la quantité de données uploadées par les appareils vers les datacenters[214].



source : Federated Learning par Min Du, 2019 (50 slides)

Compositionnalité et transformers *(2017)*

La compositionnalité est l'une des tendances clés du deep learning qui vise à lui permettre d'intégrer des éléments d'intelligence symbolique. Son principe général est de tirer parti du fait que le sens d'un système dépend de celui de ses composantes et de la manière dont elles sont assemblées (ordre, liens, temporalité, ...). Appliqué à une phrase, cela signifie que son sens dépend de celui de ses parties et de la manière dont elles sont agencées. La compositionnalité est une tentative d'intégrer de la logique symbolique dans les réseaux de neurones. Elle est notamment mise en œuvre avec les transformers dans le traitement du langage[215] ainsi que dans celui de l'image[216].

---

[213] Cette cartographie provient de The Neural Network Zoo de l'Institut Asimov, septembre 2016. Chaque type de réseau de neurone est bien décrit un par un. Ce Zoo été créé par Fjodor van Veen entre juin et novembre 2016. Il a été depuis mis à jour par Stefan Leijnen, directeur de l'institut Asimov depuis 2016.

[214] Voici quelques sources pour creuser le federated learning : Voir Federated Learning, from Research to Practice par Brendan McMahan, 2019 (83 slides), Federated Learning Privacy-Preserving Collaborative Machine Learning without Centralized Training Data par Jakub Konečný, 2018 (164 slides), Federated Learning par Min Du, 2019 (50 slides), Towards Federated Learning at Scale System Design par Keith Bonawitz et al, 2019 (15 pages), Federated Machine Learning Concept and Applications par Qiang Yang et al, 2019 (19 pages), Federated Learning: Challenges, Methods, and Future Directions par Tian Li et al, 2019 (22 pages) puis Multi-Agent Visualization for Explaining Federated Learning par Xiguang Wei et al, 2019 (3 pages).

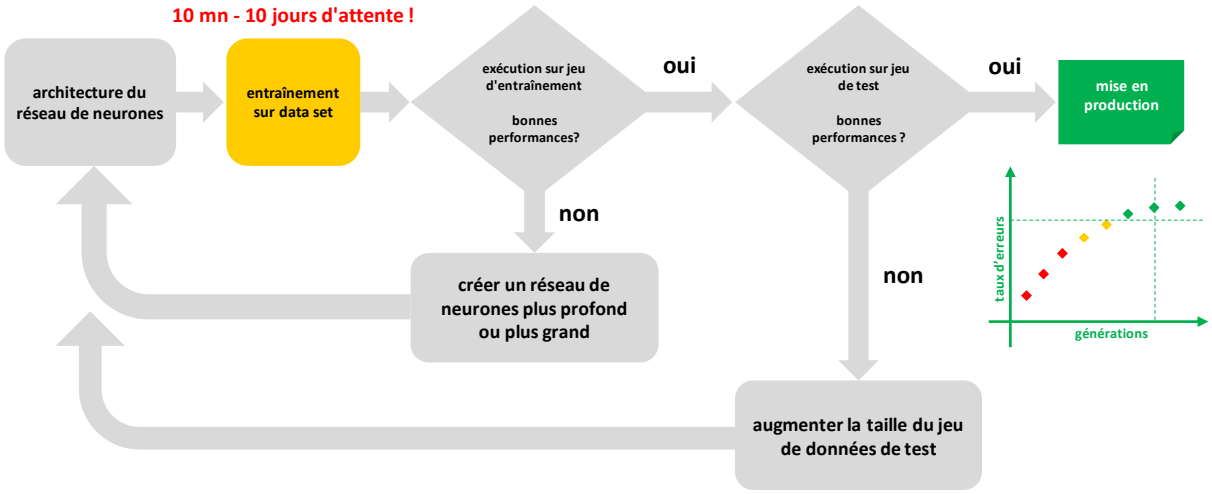
[215] Voir Towards Compositional Understanding of the World by Deep Learning par Yoshua Bengio, février 2020 (47 mn) et Building Machines That Learn and Think Like People par Brenden Lake et al, 2016 (58 pages) qui évoque les avancées récentes induites par les transformers et le besoin d'améliorer la détection de causalité pour les faire se rapprocher du raisonnement humain.

Les artifices de la compositionnalité permettent notamment de réduire le nombre d'exemples utilisés pour de l'entraînement et d'ajouter des niveaux d'abstraction dans les réseaux de neurones.

Modes d'apprentissage

Comme pour le machine learning, l'apprentissage de solutions de deep learning suit l'une des approches suivantes :

- L'**apprentissage supervisé** qui repose sur l'entraînement d'un réseau de neurones avec un jeu de données d'entraînement qui est associé à une donnée de résultat souhaité. Pour la reconnaissance d'images, il s'agit des labels ou descriptifs d'objets contenus par les images ou leur classe. Pour de la traduction automatique, ce sont des couples de phrases traduites d'une langue à l'autre. La labellisation est généralement manuelle et d'origine humaine. Tout du moins, dès lors que l'on exploite du langage. Le tagging de données peut exploiter des sources non humaines, comme des capteurs sensoriels divers.



- L'**apprentissage non supervisé** qui est utilisé dans certains types de réseaux de neurones de deep learning ou certaines parties de réseaux, comme les stacked autoencoders qui permettent d'identifier automatiquement des patterns dans des objets et de réaliser du clustering automatique d'objets. C'est un moyen de découvrir des classes d'objets nouvelles alors que dans l'apprentissage supervisé, on entraîne le modèle avec des classes préétablies manuellement. Il ne va pas identifier automatiquement le nom des classes identifiées. L'apprentissage totalement non supervisé est plus que rare. L'apprentissage non supervisé est souvent utilisé en complément d'un apprentissage supervisé.

  L'apprentissage non supervisé semble indiquer qu'une machine peut faire preuve de créativité ou d'intuition mais il n'en est rien. Il permet d'identifier automatiquement des variables discriminantes d'un élément spécifique. Dans les chatbots et le traitement du langage, l'apprentissage non supervisé est une forme d'apprentissage par renforcement. C'est le cas si le chatbot s'entraîne à mieux répondre en examinant la manière dont évoluent les conversations avec les utilisateurs en fonction de leurs réponses. Cela implique donc encore une boucle avec des humains.

- L'**apprentissage par renforcement** qui consiste à faire évoluer un modèle en fonction de retours externes, en général avec le monde physique. Cette méthode vise à optimiser une récompense quantitative (reward) obtenue par le système en fonction de ses interactions avec son environnement. C'est une technique qui est par exemple utilisée pour optimiser le réalisme des dialogues de chatbots, dans les jeux vidéo ou en robotique.

---

[216] Voir Transformers in Computer Vision par Cheng He, janvier 2021.

Elle l'est dans les robots qui apprennent à éviter les obstacles ou à réaliser des tâches mécaniques en tâtonnant. L'agent à entraîner par renforcement cherche à maximiser par itérations successives une récompense qui est incarnée par sa performance, telle que le temps pour réaliser une tâche donnée ou la qualité de cette tâche. L'apprentissage par renforcement peut nécessiter un grand nombre d'essais et de données[217].

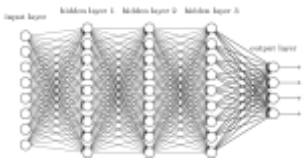
- **L'apprentissage semi-supervisé** (semi-supervised) qui exploite un mélange de données labellisées et non labellisées. Intuitivement, il s'agit d'utiliser les corrélations entre données non labellisées et non labellisées pour ajouter des labels à ces dernières.
- **L'apprentissage auto-supervisé** (self-supervised learning) est une variante de l'apprentissage supervisé qui s'appuie sur des techniques de labellisation avec des robots logiciels ou des règles logiques de représentation d'images. Elle peut exploiter des données non visuelles pour fournir des labels aux données lors de l'entraînement, trouver des corrélations entre différents types de données et ainsi labelliser des données sans contribution humaine directe.[218]
- **L'apprentissage par transfert** qui consiste à exploiter un réseau de neurones déjà entraîné pour en créer un nouveau qui va ajouter de nouvelles classes d'objets sans avoir à refaire la totalité de l'entraînement. C'est une sorte d'apprentissage incrémental.

Applications du deep learning

Depuis 2012, le deep learning est mis à toutes les sauces, la plus symbolique étant la victoire de **DeepMind** contre le champion du monde de Go à la mi-mars 2016.

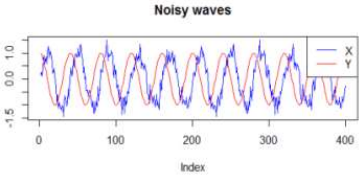
Le deep learning est surtout utilisé aujourd'hui pour la reconnaissance des formes dans les images et celle de la parole, donc dans les sens artificiels.

Il peut aussi servir à exploiter des données textuelles non structurées et à alimenter des bases de connaissances qui elles-mêmes seront exploitées par des moteurs de règles dans des systèmes experts utilisant une logique formelle ! IBM liste quelques-unes de ces applications dans **son marketing**.

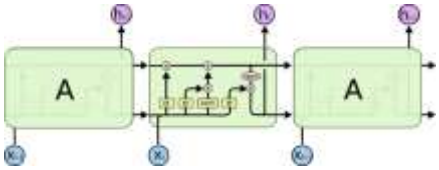


fully connected (FCN)
classification
et prédictions

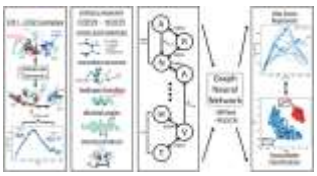
convolutifs (CNN)
spatial
reconnaissance d'images



récurrents (RNN)
temporels
ECG, finance, bruit



à mémoire (LSTM)
contexte - bidirectionnel
traduction, dialogue, recherche



à graphes (GNN)
chimie, optimisation
classification + génération

génératifs (GAN)
variations – augmentation
modification d'images et de textes

[217] Voir ce cours en 10 sessions de 1h30 en anglais sur l'apprentissage par renforcement de David Silver.

[218] Voir Self-Supervised Learning par Andrew Zisserman, Google, 2018 (122 slides).

On y retrouve des études de cas dans l'éducation pour créer des MOOC auto-adaptatifs, dans le retail avec un assistant d'achats, dans la santé avec la personnalisation de traitements contre certains cancers ou encore dans l'analyse de diverses données dans la smart city.

Pour comprendre le fonctionnement du deep learning dans le détail, il faut avoir beaucoup de temps et un bon bagage mathématique et logique[219] !

Le deep learning est très coûteux en ressources machines, surtout pendant les phases d'entraînement. Nous avons vu que celui-ci passe par l'optimisation des poids de centaines de millions de neurones qui doit être réalisée en testant chaque objet de référence en entrée et il peut y en avoir plusieurs millions. Chaque traitement d'une image de référence peut nécessiter des milliards d'opérations. Les processeurs traditionnels ne sont pas bien adaptés à ces traitements. En effet, ils vont tester et adapter séquentiellement le poids des synapses de chaque neurone et la construction des « feature maps » des couches de convolution.

Du côté du livre des records :

- En 2011, **Google Deep Brain** reconnaissait des chats dans des vidéos YouTube avec un réseau comprenant 1,7 milliards de connexions, 10 millions d'images de 200x200 pixels, 1000 machines et 16 000 cœurs, 170 serveurs, 20 000 catégories d'objets et 3 jours de calcul[220].
- En 2013, une équipe de **Stanford** sous la direction d'Andrew Ng créait un réseau de neurones de reconnaissance d'images de 11,2 milliards de paramètres tournant sur 16 serveurs à base de GPU Nvidia[221].
- En 2015, le **Lawrence Livermore Lab** créait un système gérant 15 milliards de paramètres[222] pour reconnaître des objets dans une base de 100 millions d'images issue de Flickr.
- Toujours en 2015, la startup **Digital Reasoning** de Nashville exploitait un réseau de neurones de traitement du langage cherchant des analogies parmi 20 000 mots et avec 160 milliards de paramètres, entraîné sur trois serveurs en une journée[223]. Avec un taux d'erreur de moins de 15%, un record à l'époque.
- Encore en 2015, on passait à la reconnaissance de visages avec **Nvidia**, toujours sur 100 millions d'images, avec 10 niveaux de neurones, un milliard de paramètres, 30 exaflops et 30 GPU-jours de calculs pour l'entraînement[224].
- En 2017, ce sont les **réseaux génératifs** qui ont le plus impressionné avec leurs capacités à générer des visages de synthèse, à compléter des images incomplètes ou à coloriser des photos noir et blanc.
- En 2018, **Samsung** présentait une TV 8K censée faire de l'upscaling d'images SD, HD et UHD en 8K grâce à des réseaux de neurones génératifs. On est en droit de douter mais avec une bonne base d'entraînement, pourquoi pas. Cette fonctionnalité était ensuite intégrée dans la majorité des TV 8K haut de gamme et même de certaines TV 4K lors du CES 2019.

---

[219] On peut commencer par parcourir Deep Learning in Neural Networks de Jürgen Schmidhuber, 2014 (88 pages dont 53 de bibliographie) ou bien Neural Networks and Deep Learning, un livre gratuit en ligne qui expose les principes du deep learning. Il explique notamment pourquoi l'auto-apprentissage est difficile. Il y a aussi la masse Deep Learning de Ian Goodfellow, Yoshua Bengio et Aaron Courville, de 802 pages[219]. Vous pouvez aussi visionner la conférence inaugurale de Yann Le Cun au Collège de France en février 2016 où il excelle dans la vulgarisation même si l'on peut avoir du mal à suivre jusqu'à la fin la première fois.

[220] Voir Google's artificial brain learns to find cat videos, Wired, 2012.

[221] Voir GPU-Accelerated Machine Learning and Data Mining Poised to Dramatically Improve Object, Speech, Audio, Image and Video Recognition Capabilities, Nvidia, 2013.

[222] Voir Large-scaled deep learning ont the YFCC100M dataset, 2015.

[223] Voir Biggest Neural Network Ever Pushes AI Deep Learning, et Modeling Order in Neural Word Embeddings at Scale, 2015.

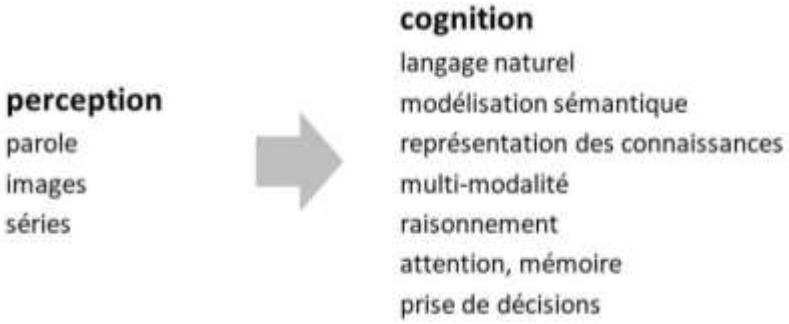
[224] Voir Deep learning image classification, Nvidia, 2016.

Ces performances vertigineuses s'expliquent notamment par la vague de l'usage de GPU et de processeurs neuromorphiques dont la structure interne est plus adaptée aux calculs des réseaux de neurones que les CPU traditionnels.

Ces processeurs savent paralléliser les calculs et multiplier des matrices entre elles, ce qui est utile pour les réseaux de neurones convolutifs. Nous verrons dans une partie suivante comment progressent ces processeurs spécialisés.

Jusqu'à présent, nous avons évoqué les applications du deep learning dans la reconnaissance des formes. Le deep learning a-t-il d'autres usages, notamment dans le cognitif et dans l'intelligence symbolique, jusqu'ici l'apanage des systèmes experts ? Oui !

Ces techniques dites cognitives sont des techniques avancées de traitement du langage avec une vision plus statistique que logique. Il reste un sacré chemin à parcourir pour intégrer de la logique et du raisonnement dans les réseaux de neurones. Cela reste un champ prolifique de recherche.



Outils du deep learning

Poursuivons cette partie sur le deep learning en évoquant l'offre des outils de création des solutions les mettant en œuvre. Il s'agit d'outils de développement exploitant des langages déclaratifs comme Python. Ils permettent de créer des modèles de réseaux de neurones avec leurs différentes couches.

La programmation consiste surtout à définir la structure du réseau de neurones : le nombre de couches cachées, la taille des filtres et des feature maps pour les réseaux de neurones convolutifs, les fonctions de pooling (réduction de résolution), puis à déclencher son entraînement avec une boucle de programme qui va scanner un jeu d'entraînement taggé et faire de la rétropropagation de gradient dans le réseau de neurones. Une fois entraîné, on évalue le taux d'erreurs généré sur un jeu de test et on affine tous les éléments ci-dessus dans ce que l'on appelle l'optimisation des hyperparamètres.

Les hyperparamètres définissent la structure d'un réseau de neurones : le nombre et la nature des couches ainsi que le nombre de filtres et leur taille dans les couches de convolution. Les paramètres sont les variables du réseau (poids des liaisons entre neurones et contenu des filtres) qui sont définies lors de l'entraînement du réseau de neurones.

Les outils disponibles pour créer des solutions de deep learning sont le plus souvent disponibles en open source, installables sur les machines et serveurs des utilisateurs ou accessibles via des ressources serveur en cloud.

Les grands acteurs du numérique proposent tous leurs frameworks et outils open source de création de réseaux de neurones : TensorFlow chez Google, PyTorch / Caffe chez Facebook, CNTK chez Microsoft, la plateforme Watson chez IBM ou encore DSSTNE chez Amazon.

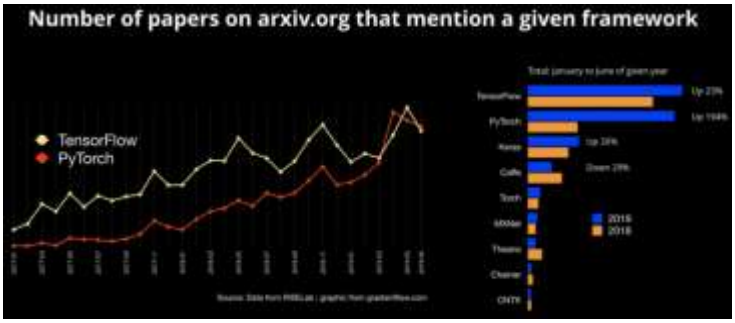
Les modèles de réseaux de neurones se définissent soit avec des fichiers de configuration (Caffe, CNTK) soit par langage de programmation et notamment Python (PyTorch, TensorFlow) ou encore Lea (pour Torch). Python est le langage le plus utilisé dans ce domaine. Ça tombe bien car il sert aussi à développer la partie back-end de nombreux sites web. Il est d'ailleurs aussi utilisé de manière standard dans la programmation d'ordinateurs quantiques.

L'un des frameworks sort du lot, tout du moins côté usage chez les startups, est **TensorFlow** dont le développement a été initialisé par Google[225]. Il fonctionne en embarqué aussi bien que sur serveurs et dans le cloud[226]. C'est un framework avec un spectre fonctionnel très large, et qui se déploie facilement sur des architectures parallèles, et notamment celles qui sont à base de GPU comme ceux de Nvidia.

Le nom TensorFlow vient de Tensor qui décrit les matrices multidimentionnelles gérées par le système. Google a annoncé au Google I/O de juin 2017 la sortie de **TensorFlow Lite**, une version allégée dédiée aux smartphones tournant sous Android. Il existe même une version **Tensorflow.js** pour les navigateurs et qui peuvent tourner sous node.js ainsi que **TFLearn**, une version simplifiée de TensorFlow pour l'apprentissage du développement.

Theano est un projet académique lancé par l'Université de Montréal qui était initialement très bien supporté et apprécié pour sa rapidité de fonctionnement. Il était très couramment utilisé par les startups jusqu'en 2016. Mais TensorFlow a pris le dessus depuis 2016 à une vitesse incroyable et l'équipe de Theano annonçait qu'il n'était plus supporté à partir de 2018. Un retournement de marché en moins de deux ans ! Cela montre le côté impitoyable des batailles de plateformes et la difficulté que peuvent avoir des laboratoires de recherche à l'origine de projets open source lorsqu'ils sont face à de grands acteurs du numérique avec leur force marketing. C'est une leçon à retenir pour les laboratoires français, comme Inria et son framework **scikit-learn** qui jusqu'ici résiste plutôt bien.

Une solution presque tout aussi populaire que TensorFlow est **PyTorch** (2016), un sur ensemble de Torch exploitable en Python. Alors que les données sont définies de manière statique dans TensorFlow, elles le sont de manière dynamique dans PyTorch, apportant une plus grande souplesse dans le développement. PyTorch est surtout utilisé par les chercheurs[227].



[225] A l'origine de TensorFlow, le projet DistBelief lancé en 2011 par Google Brain, un système de deep learning.Avec l'aide de Jeff Dean, le projet avait été réorganisé et fiabilisé. Cela a donné TensorFlow.

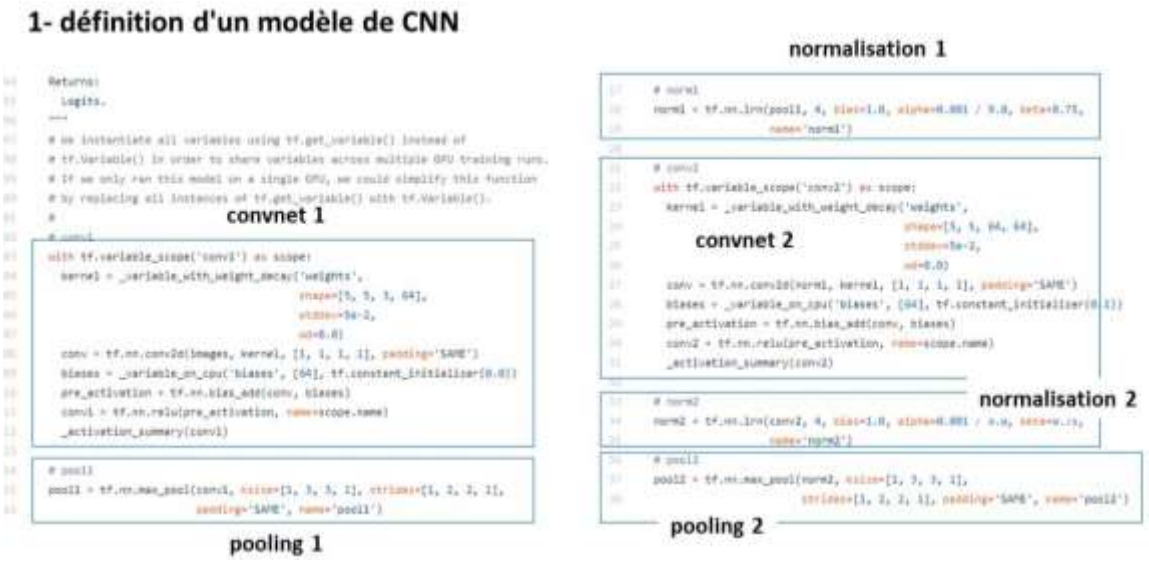
[226] Voici une source récente qui indique la popularité des frameworks d'IA : AI frameworks and hardware : who is using what?, mai 2018.

[227] Le graphe est issu de One simple graphic: Researchers love PyTorch and TensorFlow - Interest in PyTorch among researchers is growing rapidly par Ben Lorica, juillet 2019.

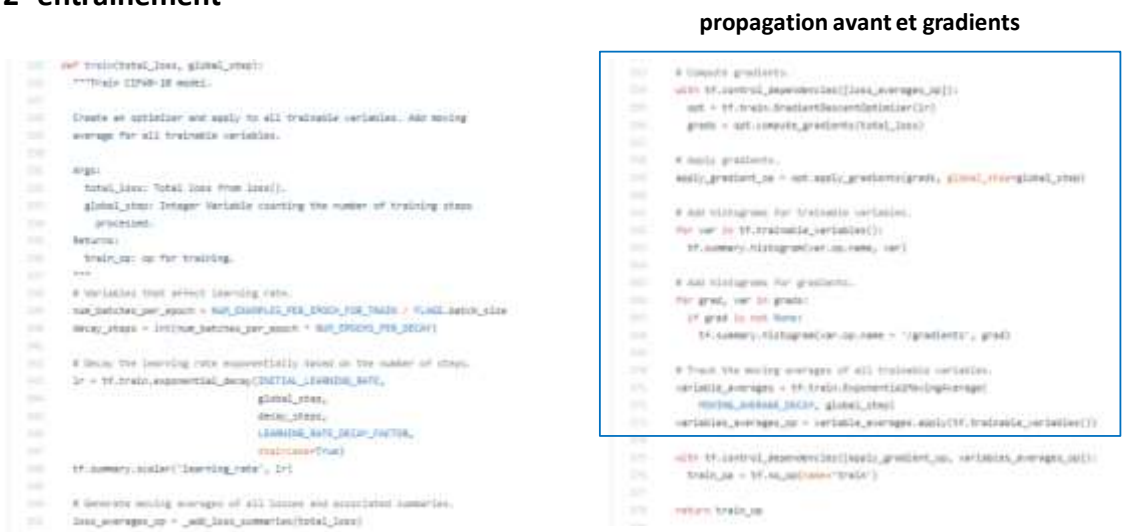
En 2020, Pytorch montait en puissance par rapport à TensorFlow en particulier auprès des chercheurs[228] mais TensorFlow reste en tête sur plusieurs indicateurs tels que les postes à pourvoir et les questions posées par les développeurs sur StackOverflow[229].

Caffe (2017) est utilisé de son côté pour les déploiements, PyTorch et Caffe provenant tous deux de Facebook. En fait, Caffe a été intégré dans PyTorch en avril 2018. La version 2 avait été créée par Yangqing Jia, un chercheur de Berkeley ensuite embauché par Facebook. Les processus de développement entre prototypage dans PyTorch et mise en production dans PyTorch sont maintenant intégrés.

D'un point de vue pratique, à haut niveau, la programmation d'un réseau de neurones de deep learning revient à définir le modèle du réseau lui-même en décrivant de manière quasiment littérale une à une toutes ses couches (*ci-dessous*, un exemple en Tensorflow et Python).



Il faut ensuite programmer son entraînement, ce qui peut requérir de l'optimisation programmatique, puis son exécution en mode run-time (*ci-dessous*, toujours avec du code Python utilisant Tensorflow).



[228] Voir PyTorch contre TensorFlow : Facebook prend le dessus sur l'IA de Google par Antoine Crochet-Damais, mai 2020. Selon PapersWithCode, 47% des articles de chercheurs publiés avec du code associé s'appuient sur PyTorch en date de septembre 2020.

[229] Voir TensorFlow Turns 5 - Five Reasons Why It Is The Most Popular ML Framework par Janakiram MSV, novembre 2020.

L'optimisation d'un réseau de neurones peut dépendre des capacités de l'architecture matérielle exploitée. Ainsi, la taille des filtres dans les réseaux convolutifs pourra être liée à celle des multiplicateurs de matrices des GPU ou processeurs neuromorphiques utilisés dans les serveurs d'entraînement.

En pratique, on n'est pas obligé de développer ses réseaux de neurones sous TensorFlow chez Google. Comme l'indique le schéma ci-contre, on peut aussi utiliser des modèles préfabriqués de machine learning qui sont personnalisables, voire même des modèles prêts à l'emploi comme pour la reconnaissance d'images ou le traitement du langage.

Certains outils sont exploités de manière combinée. Ainsi, la bibliothèque de prototypage de deep learning **Keras**, créé par le Français François Chollet (qui travaille chez Google depuis 2015), peut-elle s'appuyer sur **TensorFlow**. Ces différents outils sont aussi disponibles dans des offres en cloud, notamment chez Google, Amazon, Microsoft, IBM et même chez OVH.

Voici quelques-unes des solutions de deep learning les plus courantes pour les développeurs de solutions extraites de l'édition 2019 du Guide des Startups.

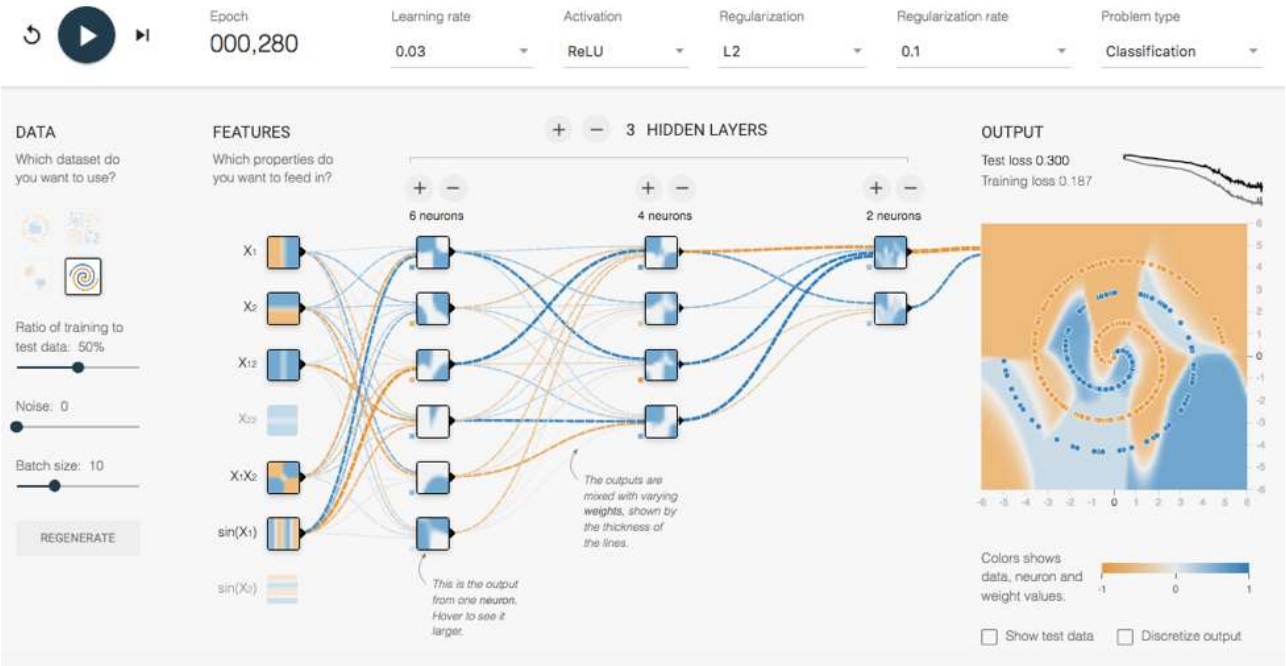
Outil	Usage
TensorFlow	**TensorFlow** est une bibliothèque open source de développement d'applications de machine learning déployable dans le cloud de manière répartie ainsi que dans l'embarqué. Elle est proposée sous forme de service en cloud par Google. Elle sert notamment à détecter des patterns, à faire de la classification automatique et des prévisions. Les Tensor Processing Units sont des processeurs dédiés au traitement avec TensorFlow qui ont été développés par Google pour son offre en cloud. Ils ont été notamment utilisés pour faire gagner DeepMind au jeu de Go en 2016. En 2019, les TPU en étaient déjà à leur troisième génération. Les applications TensorFlow sont faciles à déployer sur des architectures réparties sur plusieurs CPU et GPU. La version la plus récente, lancée en septembre 2019, est la 2.0 avec différentes améliorations comme une meilleure intégration de Keras, le support de traitements distribués, etc.
scikit learn	**scikit-learn** est un kit de développement d'applications de machine learning mettant en œuvre les méthodes de classification, de régression (prévisions) et de clustering. Il s'exploite en Python. La solution est en open source sous license BSD et est issue d'Inria et de Telecom Paritech. Sa communauté internationale comprend 1135 contributeurs depuis sa création avec environ 70 actifs par version. La documentation de scikit-learn fait 2971 pages (release 0.24.0 décembre 2020) !
K	**Keras** est une bibliothèque open source écrite en Python qui s'appuie sur Deeplearning4j, Tensorflow ou Theano. Elle permet de créer des solutions de deep learning avec un plus haut niveau d'abstraction que TensorFlow. Elle est issue du projet ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), et son principal auteur et contributeur est un français, François Chollet, qui travaille chez Google.
Microsoft CNTK	**CNTK** est un framework open source de Deep Learning de Microsoft qui fait partie de leur Cognitive Toolkit qui permet notamment de créer des agents conversationnels (chatbots). Microsoft propose une gamme d'API complète pour presque toutes les applications de machine learning et de deep learning.

	**IBM Watson** est la solution d'intelligence artificielle d'IBM. C'est en fait un ensemble de briques logicielles assez complet permettant de bâtir différentes formes d'applications d'intelligence artificielle, avec ce qui concerne la reconnaissance des images, de la parole ou de textes ainsi que l'automatisation du raisonnement avec des moteurs de règles et des solveurs. La solution a des usages multiples : création de robots conversationnels, aide au diagnostic et à la prescription dans l'imagerie médicale, prévisions dans la finance, aide juridique automatisée, cybersécurité, etc. Watson est notamment fourni en cloud. Il est assez couramment utilisé par les startups, IBM étant très actif dans leur recrutement.
	**MxNet** est une bibliothèque supportée par Amazon supportant le développement en Python, C, C++ et Julia. Elle est distribuable sur plusieurs GPU et CPU.
	**Clarifai** est une solution de deep learning en cloud qui sert notamment à la reconnaissance d'images, en particulier dans la santé et pour la création de moteurs de recherche d'images.
	**PaddlePaddle** est la bibliothèque de deep learning de Baidu adaptée aux traitements distribués. Elle est fournie avec un outil de visualisation de l'état de l'entraînement de son réseaux de neurones.
	**PyBrain** (Python-Based Reinforcement Learning, Artificial Intelligence, and Neural Network Library) est une bibliothèque création d'applications de machine learning et de réseaux de neurones bâtie en Python. Elle a été crée en 2010 par des chercheurs allemands et suisses financés par l'Union Européenne.
	**Caret** (Classification And REgression Training) est une bibliothèque permettant de développer des applications de prévisions à base de machine learning. La bibliothèque s'utilise avec le langage R. La bibliothèque semble dater d'environ 2008 et continue d'évoluer à ce jour.
	**Vowpal Wabbit** est une bibliothèque open source provenant de Yahoo! Research et gérée par Microsoft Research. Elle permet de créer des solutions de machine learning en ligne. Comme la grande majorité des bibliothèques de machine learning, elle sert à faire de la classification automatique, de la prévision, de la segmentation. Elle exploite des CPU multi cœurs et est compilée à partir de C++.
	**Caffe2** est un framework open source générique de deep learning. La première version provenait du Berkeley Vision and Learning Center et avait été développée avec l'aide financière de Nvidia et Amazon. La seconde a bénéficié de la contribution de Facebook. Le framework sait notamment exploiter les serveurs GPU Nvidia et en mode distribué. Les réseaux entraînés sont aussi facilement déployables sur architectures mobiles. On utilise Caffe en Python ou C++. Le framework facilite la distribution de traitement sur plusieurs processeurs. Caffe2 est intégré dans PyTorch depuis avril 2018.
	**Torch** est un framework de deep learning utilisé notamment dans la vision artificielle. Il est utilisé chez Facebook, Google et Twitter et provient de l'Université de New York. On l'exploite notamment avec le langage Lua qui est une sorte de Python simplifié. C'est le framework préféré de Yann Le Cun ! La déclinaison **PyTorch** qui est exploitable avec Python, un langage plus populaire. PyTorch est apprécié pour le prototypage de solutions de deep learning. PyTorch permet de générer dynamiquement les graphes des réseaux de neurones programmés au moment de l'exécution du code ce qui facilite leur mise au point.
	**spaCy** est une bibliothèque open source de traitement du langage pour Python. Elle permet d'analyser rapidement le contenu de textes en anglais, français, allemand et espagnol. Elle s'interface avec TensorFlow. Keras et scikit-learn.

	**H2O.ai** propose un framework open source de machine et deep learning couramment utilisé par les data scientists. La startup qui en est à l'origine a levé $146,1M. Elle est associée à un backend de distribution de traitements (Map/Reduce). Elle est exploitable à partir de nombreux langages comme R, Python, Java et Scala et via des API REST. Au passage, au printemps 2017, H2O, Continuum Analytics et MapD Technologies lancaient l'initiative GPU Open Analytics Initiative (GOAI) pour créer un framework ouvert commun destiné à l'exploitation en mémoire d'analytics sur GPU. Le tout avec la bénédiction de Nvidia. H2O propose aussi une solution d'Auto ML, qui permet d'identifier automatiquement le bon modèle de machine learning à appliquer sur un jeu de données structurées[230].
	Originaire de la fondation Apache, **Mahout** est un framework qui permet de développer des applications d'IA scalable, en particulier dans des applications de classification automatique et de filtrage collaboratif. Il est notamment utilisé chez Amazon. Mahout est la brique de base des architectures de big data.
	Egalement originaire de la fondation Apache, **MLlib** est une bibliothèque de machine learning pour Apache Spark, un système de répartition de traitements concurrent de Hadoop. On l'exploite en Java, Scala, Python et en R. Elle servira à traiter de gros volumes de données pour les exploiter avec des algorithmes de classification, régression ou segmentation automatique.
	**OpenNN** est une classe open source en C++ de la startup espagnole Artelnics qui sert à développer des applications de réseaux neuronaux à bas niveau. Les frameworks de type TensorFlow permettent de s'en passer.
	PredictionIO est une solution open source pour développer des applications de machine learning dédiées à la prévision. C'est une sorte de "MySQL pour le machine learning". C'est encore une solution de la fondation Apache. A ne pas confondre avec l'offre de la startup française Prevision.io qui propose une solution d'AutoML.
	**Algorithmia** est une place de marché d'algorithmes et de briques logicielles d'IA qui sont positionnées comme des « micro services », et disponibles en cloud, faciles à tester, intégrer et mettre en production. Les services proposés sont assez classiques comme la détection de visage dans des photos ou l'analyse de sentiments dans les flux de réseaux sociaux.
	Stanford **CoreNLP** est un framework généraliste de traitement du langage qui permet d'identifier la structure grammaticale et syntactique de phrases, de détecter des mots clés, des nombres, analyse des sentiments, etc. Il fonctionne dans de nombreux langages. Les fonctions sont exploitables via les principaux langages du marché ainsi qu'au travers de services web. Le framework est open source.
	**Open NLP** est un framework de traitement du langage de la fondation Apache créé en 2004 et qui a évidemment bien évolué depuis. Il comprend les fonctions classiques de décomposition de phrases en entités, d'analyse de structure grammaticale et syntaxique et d'extraction de données.

Vous pouvez aussi simuler des réseaux de neurones simples avec cet outil exploitant **TensorFlow** (exemple *ci-dessous*). TensorFlow peut en effet servir aussi bien à gérer un réseau de neurones multicouches de deep learning tout comme des solutions plus simples de machine learning, à base ou pas de réseaux de neurones simples.

[230] Voir The different flavors of AutoML, Erin LeDell, avril 2019.



Vous pouvez télécharger ces deux ouvrages : Deep Learning in Python de François Cholet 2018 (386 pages) et Deep learning with Keras 2017 (490 pages). Et aussi reproduire chez vous ces 30 tests d'applications de deep learning qui sont en open source[231].

Citons une dernière catégorie d'outils, les places de marché de modèles d'IA pour les développeurs. En voici quelques-unes :

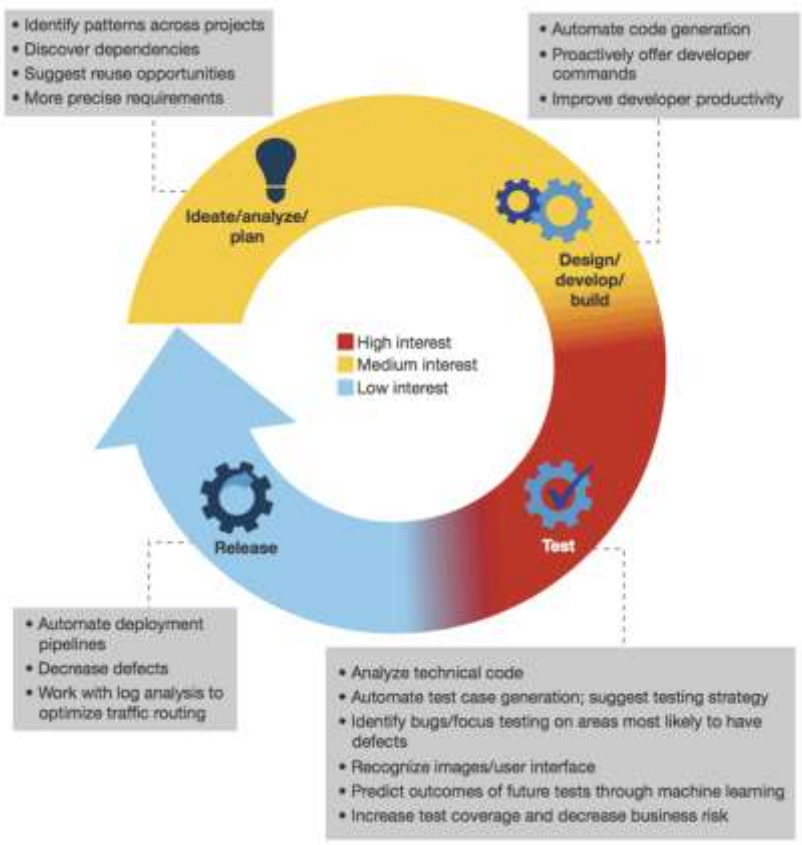
- **ONNX** (2016, USA) est un format de modèles de deep learning et machine learning interopérable supporté par Caffe2, Chainer, MxNet, PyTorch, PaddlePaddle, Microsoft mais pas TensorFlow. Peut-être parce que c'est un projet communautaire poussé par Amazon, Facebook et Microsoft. La solution permet l'import et l'export de modèles de et vers ces différents outils.
- **SingularityNET** (2017, Suisse, $36M) est une plateforme de monétisation de services d'IA. C'est une sorte de version commerciale d'ONNX. Elle a été cofondée par le fameux Ben Goertzel, un spécialiste de l'AGI. Mais là, ce n'est pas encore de l'AGI[232]. Je décris cette initiative dans la partie liée à l'AGI.
- **MLIR** (Multi-Level Intermediate Representation) est vu de loin un équivalent d'ONNX pour Tensorflow exploitable par d'autres frameworks de haut niveau. Il a été lancé par Google en 2019. C'est un langage intermédiaire entre les frameworks de machine learning et les différentes architectures matérielles cibles. La compilation peut être parallélisée sur plusieurs CPU[233].
- **Algorithmia** (2013, USA, $38,1M) est autre place de marché d'algorithmes d'IA. Elle propose à ce jour 5000 algorithmes, probablement pas tous du domaine de l'IA.

[231] Voir 30 Amazing Machine Learning Projects for the Past Year (v.2018), janvier 2018.

[232] Voir le livre blanc SingularityNET: A decentralized, open market and inter-network for AIs, décembre 2017 (53 pages) et Toward Grand Unified AGI.

[233] Voir MLIR: A new intermediate representation and compiler framework, avril 2019 et la présentation MLIR: Multi-Level Intermediate Representation Compiler Infrastructure, avril 2019 (71 slides).

Le machine learning et le deep learning sont aussi mis à contribution pour améliorer la productivité des développeurs d'applications de toutes sortes. Certains sont mis en œuvre via des solutions en cloud[234].



Ces outils à base de machine learning assistent la saisie et l'édition de code avec des fonctions qui devinent la fin d'un bloc ou d'une ligne de code en cours de saisie (autofill), le réusinage de code (code refactoring), fournissent des outils d'analyse et de gestion des erreurs, de gestion de projet, de gestion des risques et aussi de debug[235].

En voici quelques-uns :

- **DeepCode** (2016, Suisse, $5,2M) qui analyse le code pour l'améliorer à partir d'un corpus de 250 000 règles de bonne programmation. Il peut consulter le code que vous avez publié sur Github et en fait un audit de qualité et de détection de bugs potentiels. Ils proposent aussi JSNice, qui *déobsfuscate* du code JavaScript minifié. C'est une spin-off de l'Université ETH Zurich[236].
- **TabNine** (USA) propose une fonction qui aide les développeurs à coder en remplissant automatiquement les lignes de code. Une sorte d'autocomplete, à base de deep learning. Il est notamment disponible sous la forme d'un plugin de Visual Studio. La startup a été acquise par Codota en 2019.
- **Functionize** (2015, USA, $19,12M) propose un outil de tests d'applications utilisateurs permettant d'exprimer un parcours visiteur en langage naturel, qui est ensuite analysé par l'outil et exécuté via un plugin tournant dans le navigateur Chrome. L'outil génère aussi des indicateurs de performance du code.
- **Applitools** (2013, Israël, $41,8M) réalise comme Functionize des tests fonctionnels d'applications web. Il utilise de la reconnaissance d'image pour analyser les écrans et identifier des variations ou régressions en ne tenant compte que des éléments perceptibles par l'œil humain.
- **Testim** (2014, USA, $15,6M) propose un outil de gestion de tests de logiciels. Il s'appuie sur du machine learning pour accélérer la création, l'exécution et la maintenance de tests automatisés.

---

[234] Schéma provenant de 6 ways AI transform software development par Mariya Yao, mars 2018.

[235] Voir 9 Ways To Implement Artificial Intelligence and Agile-Powered Management in Software Development par Chandresh Patel, février 2018 ainsi que 5 ways artificial intelligence is upgrading software engineering par Melisha Dsouza, septembre 2018.

[236] Voir DeepCode supporte la révision des langages C et C++ par Serdar Yegulalp, mars 2020.

- **Sealights** (2015, Israël, $19,6M) utilise du machine learning pour analyser le code à tester et les tests qui y sont appliqués, qu'il s'agisse de tests fonctionnels ou de performance.
- **Test.AI** (2015, USA, $17,6M) utilise un outil de détection automatique des composantes visuelles d'une application pour générer des tests. Cela semble surtout cibler les applications mobiles.
- **Weights and Biases** (2018, USA, $65M) propose des outils d'optimisation d'hyperparamètres de modèles de machine learning, un gestionnaire de versions de modèles et suivi d'expérimentation.
- **Mabl** (2017, USA, $36,1M) utilise le machine learning pour gérer les tests et identifier les régressions dans les évolutions du code. L'outil adapte automatiquement les tests en fonction de la détection d'évolutions de l'interface utilisateur de l'application. La société a été créée par des anciens de Google.
- **Diffblue** (2016, UK, £17M) propose un générateur de tests unitaires pour du code réalisé en Java. Il nécessite 64 Go de mémoire et tourne sous Linux (Ubuntu, RedHad ou CentOS). Il s'appuie sur du machine learning pour automatiser ou assister les tâches de correction de bugs, de trous de sécurité, de conversion entre langages de programmation ou de création de code. C'est le résultat d'une dizaine d'années de recherche à l'Université d'Oxford.
- **Ponicode** (France, 2019, 4M€) propose un outil à base de machine learning qui génère automatiquement des tests unitaires de code et propose des améliorations aux développeurs. Il avait lancé sa bêta en juin 2020. La startup avait rejoint en mai 2020 le programme Microsoft AI Factory à Station F. La startup a été cofondée par Patrick Joubert, qui avait notamment lancé Recast.AI en 2015, acquis par SAP en 2018.
- **DarwinAI** (2017, Canada, $2M) propose une solution un peu mystérieuse qui automatise la création de solutions de deep learning en fonction de besoins (qui doivent être assez classiques et identifiés) qui sont par ailleurs explicables. Leur technologie Generative Synthesis est issue de l'Université de Waterloo. [237]
- **Miriya.AI** (2015, France) propose des générateurs de code automatiques exploitant du deep learning de traitement du langage, avec un générateur de tests unitaires de logiciels écrits en Java, C, C++ et Python ainsi qu'un générateur de projets Android qui interprète des spécifications textuelles structurées (vidéo).
- Un projet de recherche de 2020 utilise un réseau génératif GAN pour **traduire du code** d'un langage en un autre langage de programmation, plus efficacement qu'un système à base de règles[238].

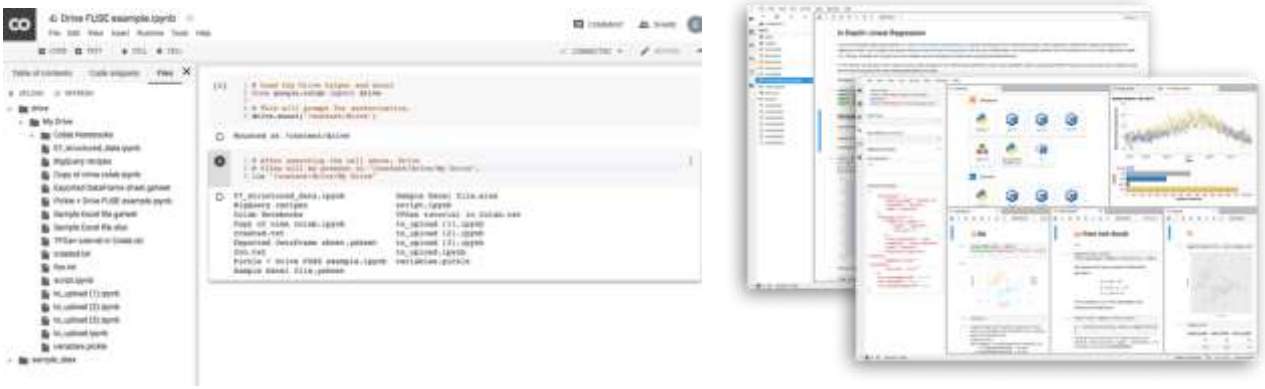
Va-t-on pour autant voir apparaître des outils de développement à base d'IA qui vont remplacer les développeurs ? C'est une thèse partagée par quelques prospectivistes de l'IA. Je n'y souscris pas. Créer un logiciel nécessite toujours de formaliser le besoin. Que ce soit de manière graphique ou par du code, on doit toujours en passer par là. Ce qui pourra arriver est la prolifération de solutions permettant d'exploiter des briques d'IA à haut niveau sans avoir à connaître les dessous du machine learning et du deep learning. Des évolutions des outils de développement et de tests à base d'IA feront cependant leur apparition, qui assisteront les développeurs. Mais pas de là à les remplacer.

[237] Voir les travaux liés à cette startup : Voir AttoNets: Compact and Efficient Deep Neural Networks for the Edge via Human-Machine Collaborative Design par Alexander Wong & al, 2019 (15 pages), NetScore: Towards Universal Metrics for Large-scale Performance Analysis of Deep Neural Networks for Practical On-Device Edge Usage par Alexander Wong, juin 2018 (9 pages) et EdgeSegNet: A Compact Network for Semantic Segmentation par Alexander Wong & Al, mai 2019 (5 pages).

[238] Voir Unsupervised Translation of Programming Languages par Marie-Anne Lachaux et al, 2020 (21 pages).

Il faut aussi citer **Jupyter Notebook** (vidéo) et **Google Colab** (vidéo) qui sont des outils de développement collaboratifs de modèles de machine learning. Jupyter Notebook est un logiciel open source qui fonctionne dans un navigateur web, destiné aux développeurs et au data scientists. Il permet le partage de code et de documents divers liés au projet de développement pour le travail collaboratif.

On programme avec en Python et on accède à des bibliothèques de code plus à des outils de visualisation de données. **Google Colab** est un sur ensemble de Jupyter Notebook qui s'intègre au cloud et aux outils de travail collaboratif de Google. Il exploite notamment les processeurs TPU de Google[239].



Anaconda est un autre outil populaire pour le développement d'applications de machine learning en Python ou R, qui simplifie la gestion des paquets et leur déploiement.

Anaconda comprend plus de 250 paquets dont une bonne palanquée adaptée au machine learning, tournant sur Windows, Linux et MacOS.

Anaconda est complété par Anaconda Navigator, un environnement de développement graphique.

Le deep learning peut être aussi associé à des outils de simulation dans des approches hybrides. C'est notamment l'objet du projet collaboratif **HSA** (Hybridation Simulation Apprentissage) lancé par l'IRT SystemX en février 2020 et devant durer quatre ans. Il associe Airbus, Air Liquide, EDF, RTE et SNCF et les laboratoires de recherche Tau d'Inria Saclay, le LIP6 de Sorbonne Université, et le CNRS LIMSI. Les outils de simulation modélisent des systèmes à partir des modèles mathématiques en définissant le fonctionnement (équations aux dérivées partielles, modélisation par éléments finis, etc) tandis que les données récupérées par l'observation et des capteurs alimentent des modèles de machine learning. L'hybridation des deux peut servir à réduire le coût de la simulation et à améliorer les capacités de prédiction des modèles de machine learning. Parmi les applications évoquées : Airbus veut simuler le bruit des aéronefs pour réduire des nuisances sonores, Air Liquide veut optimiser la production d'hydrogène par reformage du méthane à vapeur et EDF optimise la modélisation énergétique des bâtiments et des quartiers.

[239] Voir All About Using Jupyter Notebooks and Google Colab par Divya Singh, mars 2019.

IA explicable

L'une des craintes associées à l'IA connexionniste en général et au deep learning en particulier est de générer des solutions logicielles dont le fonctionnement n'est pas facilement explicable, interprétable et auditable.

Les exemples les plus faciles à visualiser sont ces réseaux de neurones convolutifs qui sont capables de caractériser des images. Ils s'appuient sur des filtres dans plusieurs couches de convolution qui ne sont pas créés par l'Homme mais automatiquement générés par le mécanisme de la rétropropagation pendant l'entraînement. Ces filtres correspondent à une analyse probabiliste automatique permettant au réseau de neurones de distinguer des formes de niveaux d'abstraction empilés. Mais ces niveaux d'abstraction ne correspondent pas forcément à ceux de la vision humaine. Donc, on a du mal à comprendre le cheminement de l'information lorsque l'on traverse les différentes couches du réseau de neurones.

Il en va de même pour l'analyse du langage pour détecter des sentiments, réaliser des traductions ou extraire des données. Les modèles probabilistes du deep learning des réseaux de neurones de traitement du langage, mêlant souvent des convnets, des RNN (réseaux récurrents) et des LSTM (réseaux à mémoire), étant générés automatiquement, ils sont difficiles à auditer.

Nombreuses sont les histoires d'effets de bords indésirables de l'IA dans la prise de décision. Un exemple connu est cet ouvrier de chantier palestinien arrêté par erreur en 2017 par la police israélienne en 2017 parce qu'il avait posté une photo de lui, penché sur un bulldozer.

La photo avait été labellisée par Facebook avec un « *attaquez les* » qui était une mauvaise traduction de « *bonjour* » (good morning) en arabe. Il existe de nombreuses histoires du même genre où des systèmes à base de machine learning entraînés avec des données biaisées sont utilisés pour priver de liberté des gens, notamment aux USA[240]. Les risques de ce genre sont élevés avec les systèmes d'analyse de risques pour des prêts dans les banques ou des couvertures pour les assurances. Sans compter les systèmes de santé.

Il y a bien des cas bien pratiques où une explication serait la bienvenue. Ainsi, j'ai voulu poster un commentaire sur un produit défectueux acheté sur Amazon en août 2018.

Amazon m'a répondu que mon commentaire d'une ligne n'était pas conforme aux règles en vigueur chez eux, en oubliant de préciser ce qui n'allait pas dans mon texte et la règle violée. Ce n'était pas du tout évident en consultant lesdites règles qui sont étalées sur plusieurs pages. Même s'il est probable que le « *it's a joke* » n'était pas acceptable pour eux[241], ce qui est pour le moins étonnant.

La notion d'IA explicable est devenue un champ scientifique à part entière de l'IA[242]. On l'appelle **XAI** pour Explainable AI[243]. C'est devenu une priorité presque partout dans le monde, dans la recherche aux USA, en Chine et en France.

---

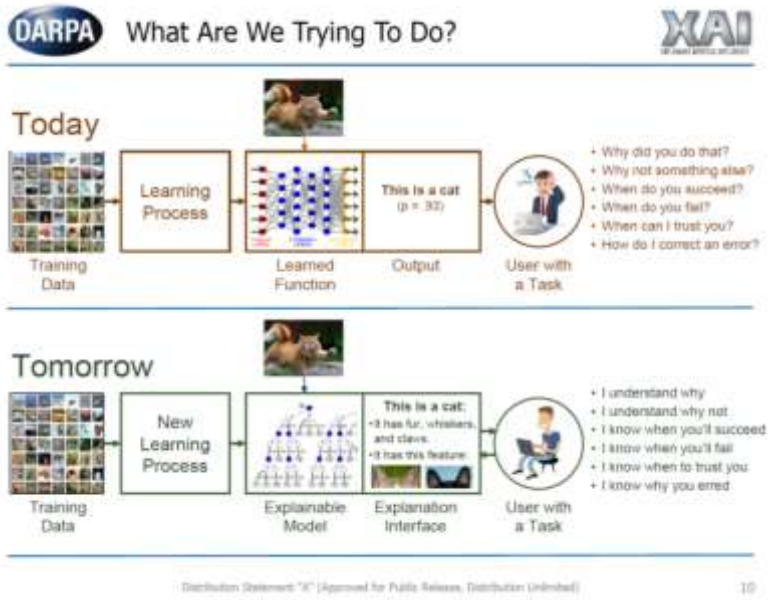
[240] Voir When a Computer Program Keeps You in Jail par Rebecca Wexler, New York Times, 2017.

[241] C'est l'objet de travaux comme Understanding Hidden Memories of Recurrent Neural Networks, 2017 (16 pages) qui vise à améliorer l'explicabilité des réseaux de neurones récurrents dans le traitement du langage.

[242] Voir Pour y voir plus clair sur les notions de transparence et d'explicabilité en IA par Aymeric Poulain Maubant, 2020, qui positionne bien les éléments de débat entre IA éthique et IA explicable.

[243] Voir les actes du workshop sur l'XAI de la conférence IJCAI de 2017, l'une des plus importantes de l'IA au monde, dans IJCAI-17 Workshop on Explainable AI (XAI), 2017 (68 pages). Et cet article de vulgarisation sur la question : The Dark Secret at the Heart of AI de Will Knight, MIT Technology Review, avril 2017.

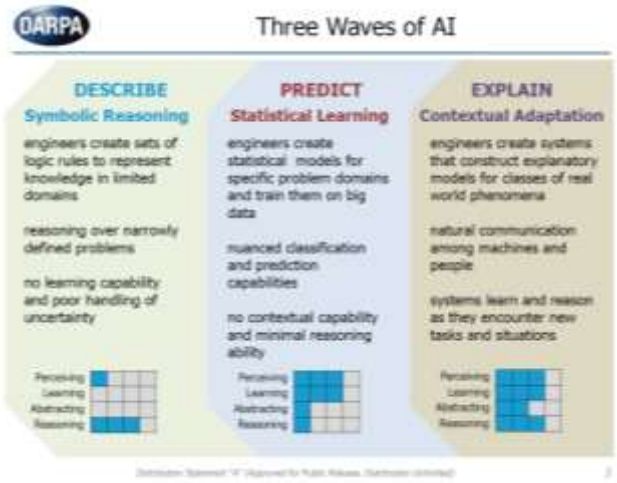
L'agence DARPA du Pentagone a lancé sa propre initiative **XAI** dédiée en 2016[244]. Dans le schéma *ci-contre*, on comprend bien l'objectif : être capable de décomposer les étapes qui mènent un réseau convolutif à reconnaître un objet (ici, un chat) dans une image.

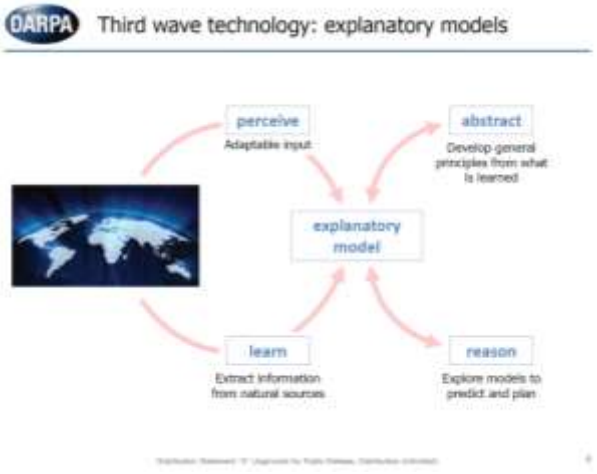


L'intention est également de mieux comprendre les biais éventuels des réseaux de neurones pour identifier les sources d'erreurs potentielles. Il faut noter une distinction entre l'explicabilité et l'interprétabilité.

L'**explicabilité** concerne en gros le processus et l'algorithme en général et l'**interprétabilité** décrit le lien entre les solutions et les données d'entraînement, à savoir quelles sont les données qui ont influencé le résultat fourni par une IA connexionniste. L'explicabilité des IA symboliques est plus facile car on peut chaîner les faits et règles qui ont été utilisés dans les raisonnements et prises de décisions[245].

Cette dernière peut s'appuyer sur la décomposition d'un système, la capacité à le simuler (donc à reproduire une éventuelle erreur) et sur la transparence des algorithmes utilisés. Et la notion d'interoprétabilité est évidemment dépendante de la personne qui peut interpréter l'interprétation ! On est à la frontière entre l'IA et la philosophie[246] !





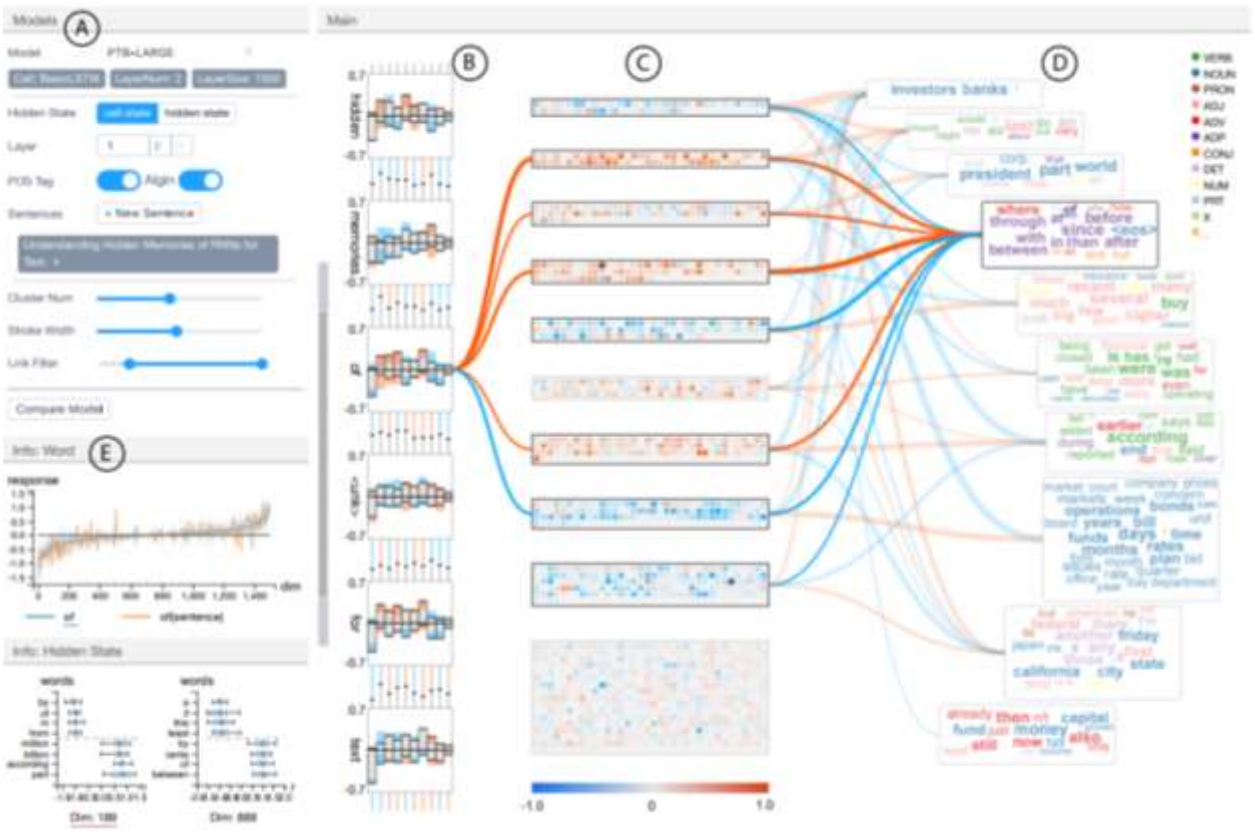
[244] David Gunning de la DARPA explique très bien toute la démarche dans Explainable Artificial Intelligence (XAI) Program Update, novembre 2017 (36 slides) et dans une autre présentation plus récente réalisée pour la NASA (38 slides). Cette initiative a sélectionné et financé divers projets dans des universités telles que Berkeley, UCLA, Oregon State et Carnegie Mellon ainsi que dans des laboratoires de recherche privés comme le PARC de Xerox, SRI International et Raytheon BBN. Voir un état des lieux dans Inside DARPA's effort to create explainable artificial intelligence par Ben Dickson, janvier 2019. En 2018, quatre projets lancés par la DARPA avaient déjà aboutit. Voir la somme Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications and Bibliography for Explainable AI, février 2019 (204 pages), qui fait un inventaire étourdissant de la production sur le thème de l'IA explicable.

[245] Voir Principles and Practice of Explainable Machine Learning par Vaishak Belle et Ioannis Papantonis, 2020 (33 pages).

[246] Voir Explaining Explanations in AI par Brent Mittelstatd et al, 2019 (10 pages).

Les questions que l'utilisateur peut se poser concernant le résultat d'une IA sont : pourquoi ce résultat a-t-il été obtenu ? Pourquoi une autre méthode n'a-t-elle pas été mise en œuvre ? Dans quelles conditions la méthode retenue fonctionne-t-elle ? Quand échoue-t-elle ? Dans quelles conditions puis-je lui faire confiance ? Comment savoir si une erreur a été commise et comment la corriger ? On devrait aussi ajouter : quelles sont les données qui ont servi à entraîner le modèle de l'IA [247]? Couvraient-elles bien l'espace du possible ? Avaient-elles des biais statistiques éventuels ? Et ces données étaient-elles adaptées au problème ?

Les méthodes rendant les IA explicables sont aussi nombreuses que les méthodes d'IA elles-mêmes[248]. Dans nombre de cas de figure, ce sont souvent des questions plus académiques que pratiques. Elles sont surtout pertinentes pour gérer les cas les plus épineux comme un audit après un accident mettant en œuvre un véhicule à conduite autonome ou l'aide à la prise de décision dans les questions médicales, surtout de vie ou de mort. Les juristes sont évidemment très sensibles à ces questions pour à la fois faire évoluer le droit et leurs pratiques[249].



Nombre de logiciels prennent des décisions pour nous sans que l'on puisse les auditer, même lorsqu'ils n'utilisent pas forcément de l'IA. Les exemples classiques concernent le ciblage publicitaire, les flux d'information dans les réseaux sociaux ou la décision de mener un contrôle fiscal sur vous ou votre entreprise !

---

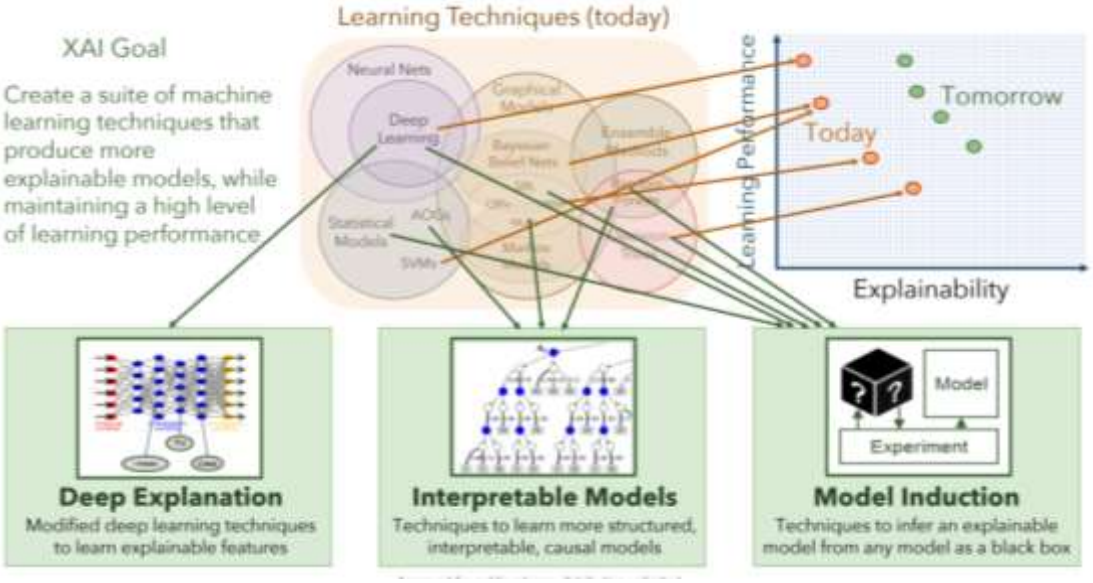
[247] Dans Increasing Trust in AI Services through Supplier's Declarations of Conformity, août 2018 (29 pages), une brochette de chercheurs d'IBM Research propose la création d'un cahier des charges d'IA explicables. Ils ajoutent aux questions de base les questions suivantes : quel est l'usage du service ? Quels algorithmes et techniques sont employés ? Décrire les méthodes de tests et leurs résultats. Existe-t-il des biais connus ayant une incidence éthique ou en matière de sécurité ? Quels efforts ont été menés pour les éviter ? Est-ce que le service a été vérifié contre les attaques de GANs (Generative Adversarial Networks) ? Quand le modèle a été mis à jour la dernière fois ?

[248] Pour ce qui est de la reconnaissance d'images, quelques méthodes de XAI sont très bien documentées dans Towards explainable Deep Learning de Wojciech Samek du Fraunhofer HHI, novembre 2017 (88 slides).

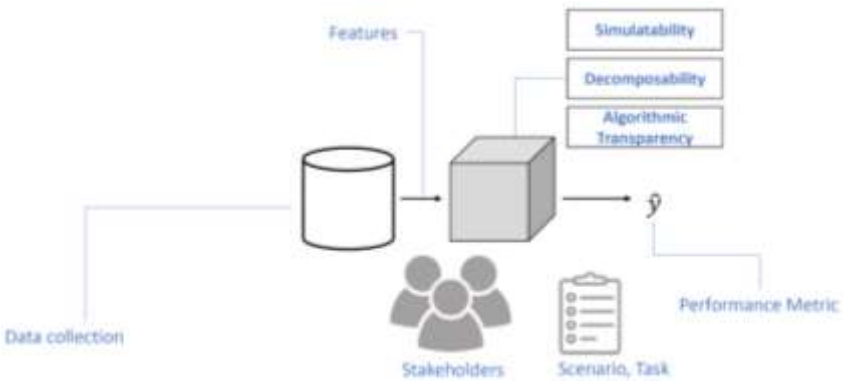
[249] Voir Accountability of AI Under the Law: The Role of Explanation, de Finale Doshi-Velez et Mason Kortz, 2017.

Néanmoins, une IA non explicable utilisée seule n'est pas compatible avec le RGPD[250]. Il ne suffit pas de pouvoir expliquer le résultat d'un traitement, il faut pouvoir en visualiser l'explication de manière compréhensible pour les utilisateurs. Cela peut relever d'approches visuelles ou d'explications en langage naturel. Cela peut aller jusqu'à la création de chatbots expliquant un résultat d'une IA.

L'un des défis des IA explicables est de bien équilibrer la recherche d'explicabilité et la performance des techniques utilisées. Comme présenté dans le schéma *ci-dessous*, le paradoxe du deep learning est d'être très performant mais faiblement explicable tandis que d'autres méthodes sont au contraire explicables mais moins performantes, comme tout ce qui relève de moteurs de règles et de logique formelle. L'idée est de conserver la bonne performance du deep learning tout en améliorant son explicabilité. La DARPA structure l'approche en trois parties : la création de modèles de deep learning explicables avec des couches intermédiaires utilisant des formes labellisables, des modèles interprétables avec des notions de causes à effets et enfin, des modèles inductifs hiérarchiques.



Il existe en fait un très grand nombre de méthodes de création d'IA explicables, sachant que des outils existent pour les nombreuses méthodes de machine learning et de deep learning. Il ne s'agit pas d'IA explicable générique[251]. Le coût de l'explicabilité peut être élevé et se faire au détriment de la qualité de l'IA.



[250] Voir GDPR / RGPD - Protection des données incompatible avec Deep Learning de Dimitri Carbonnelle Livosphere 2017. Selon son article 63 « *toute personne concernée devrait avoir le droit de connaître et de se faire communiquer [...] les finalités du traitement des données à caractère personnel, [...] la logique qui sous-tend leur éventuel traitement automatisé et les conséquences que ce traitement pourrait avoir, au moins en cas de profilage.* ». Une banque ne peut ainsi pas utiliser une solution à base de deep learning pour évaluer la solvabilité d'un client potentiel et lui refuser un prêt sans autre explication. Elle peut exploiter la solution mais doit expliquer au client pourquoi elle refuse ce prêt. Le deep learning est un outil d'aide à la décision et pas un outil de prise de décision autonome.

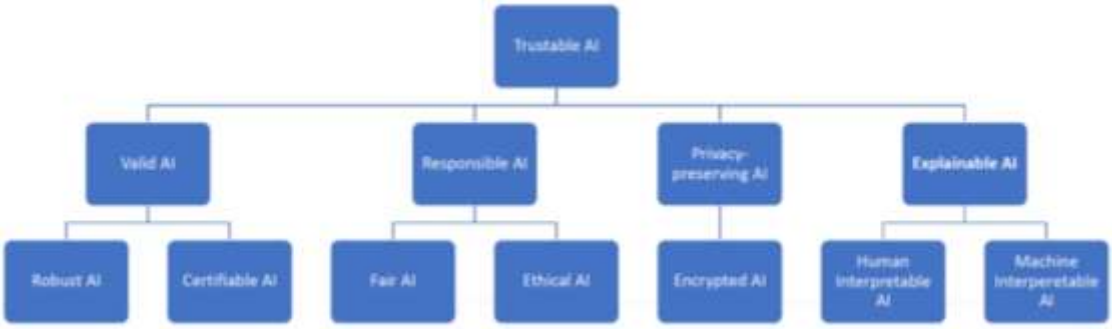
[251] Voir cet excellent AAAI 2019 Tutorial on On Explainable AI: From Theory to Motivation, Applications and Limitations, janvier 2019 qui contient la présentation On Explainable AI: From Theory to Motivation, Applications and Limitations, janvier 2019 (160 pages).

On réservera donc en général les IA explicables aux cas d'usages qui le justifient : plus dans la police et la santé que dans la recommandation de films sur Netflix, par exemple.

Idéalement, il faudrait pouvoir dialoguer via un chatbot avec une IA pour l'interroger et obtenir des éclaircissements sur ses reconnaissances ou décisions[252].

Figure 7: An example of an interactive explanatory dialog for gaining insight into a DOG/FISH image classifier. (For illustration, the questions and answers are shown in English language text, but our use of a 'dialog' is for illustration only. An interactive GUI, e.g., building on the ideas of Krause et al. [20], would likely be a better realization.)

L'IA explicable doit aussi répondre à des exigences d'éthique (biais…), de confiance (équité, rigueur, …) et de sécurité (respect de la vie privée, non détournement des données personnelles pour des usages non documentés, …)[253]. L'IA explicable est une des composantes de l'IA de confiance.



Comme l'IA washing est courante, on risque de voir émerger de l'XAI washing ou tout du moins un marketing de l'offre mettant bien l'accent sur l'explicabilité de l'IA. Cela commence avec **Maathics** (France), une startup qui propose un système d'audit automatisé d'algorithmes d'IA permettant « *un usage équitable de l'IA* ». Reste à savoir comment cela fonctionne et si le système est lui-même explicable !

Quelques avancées récentes servent à outiller les créateurs de solutions d'IA pour créer des modèles explicables.

Google What-If Tool est un outil de visualisation du comportement des modèles. Les modèles doivent être déployés sur l'AI Platform de Google dans le cloud. Il visualise les jeux de données d'entraînement de manière graphique, pour peu que le nombre de dimensions le permette et les prévisions de classification des modèles. L'outil permet aussi de comparer l'efficacité de modèles développés avec les différents frameworks supportés. Cela a l'air de bien fonctionner pour du machine learning de base sur des données multidimensionnelles, mais pas pour des réseaux de neurones de deep learning comme pour la reconnaissance d'images[254].

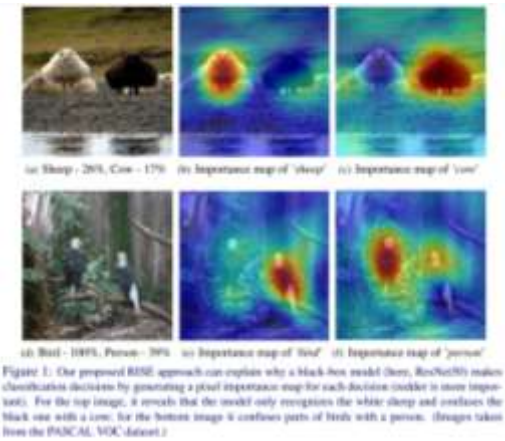
---

[252] Voir The Challenge of Crafting Intelligible Intelligence par Daniel Weld et Gagan Bansal, 2018 (8 pages).

[253] Voir Vers une intelligence artificielle ubiquitaire explicable du chercheur en cybersécurité Thierry Berthier, juillet 2018 ainsi que Explanation in Artificial Intelligence: Insights from the Social Sciences, de Tim Miller, 2018 (66 pages).

[254] Voir Google's What-If Tool And The Future Of Explainable AI par Kalev Leetaru dans Forbes, août 2019 qui fait référence à Introducing the What-If Tool for Cloud AI Platform models par Google, juillet 2019 qui fonctionne avec les modèles créés pour TensorFlow mais aussi pour Scikit Learn et XGBoost. Voir aussi A Walkthrough of the What-If Tool qui décrit bien le fonctionnement de l'outil.

L'un des biais des réseaux de neurones convolutifs se produit lorsqu'ils détectent des objets en analysant leur contexte et leur environnement plutôt qu'à partir de leur contenu. Cela amène à des erreurs de reconnaissance[255]. D'où la méthode **RISE** qui consiste à créer des « heat map » des objets reconnus dans une image pour éviter le syndrome d'erreur de contexte[256]. La méthode **SyncNet** consiste à générer des contraintes sur les formes des filtres générés lors de l'entraînement de réseaux de neurones convolutifs, appliqués à la reconnaissance de phonèmes dans la parole. Cela limite donc le choix des filtres à des filtres que l'on peut labelliser et interpréter.



Cela ne s'applique cependant pas encore à la reconnaissance d'images[257]. Il existe aussi des méthodes permettant d'expliquer le processus de certains réseaux de neurones génératifs d'images composites[258].

En juin 2019, **Thales** faisait l'acquisition de **Psibernetix** (USA), une société spécialisée dans la création d'IA certifiables, utilisée notamment dans pour son système Alpha de simulation de combat aérien. Ils utilisent un panaché de logique floue et d'algorithmes génétiques dans Genetic Fuzzy Tree ce qui permet de décrire les prises de décision.

Arthur.ai (2018, USA, $18,3M) propose un outil de suivi de l'efficacité opérationnelle de ses modèles de machine learning. Il détecte la performance des modèles, leurs déviations, les biais et l'impact de mauvaises données d'entraînement. L'outil est surtout exploité dans les applications du machine learning de suivi de la relation client. Ils travaillent aussi sur l'explicabilité du machine learning, notamment via la notion d'explication contrafactuelle. En gros, pour trouver sur une classification (refut d'un prêt pour un client) lesquels des paramètres actionnables permettraient le plus facilement de modifier la recommandation[259].

---

[255] Voir Solving AI's 'black box' problem: Learning how algorithms make decisions par Ben Dickson, novembre 2018.

[256] Voir RISE: Randomized Input Sampling for Explanation of Black-box Models par Kate Saenko & al, juin 2018 (17 pages) et Explainable AI: Interpreting the neuron soup of deep learning par Ben Dickson, octobre 2018.

[257] Voir Interpretable Convolutional Filters with SincNet par Mirco Ravanelli et Yoshua Bengio, novembre 2018 (12 pages).

[258] Voir GAN Dissection: Visualizing and Understanding Generative Adversarial Networks, 2019 qui explique le fonctionnement d'un réseau de neurones génératif de peinture (vidéo).

[259] Voir An Overview of Counterfactual Explainability, décembre 2020.

On aimerait bien qu'une IA explicable fasse son entrée dans les services achats des entreprises pour expliquer leurs lenteurs et leurs pertes de factures à répétition. Mais ils seront probablement les derniers équipés dans les grandes entreprises, tout du moins dans l'optique de payer plus rapidement les fournisseurs.

D'autres scientifiques proposent d'analyser le comportement des IA en exploitant des sciences sociales comme avec les homo-sapiens[260].

Agents et réseaux d'agents

Dans ce concept apparu dans les années 1990, les agents intelligents permettent de résoudre des problèmes dans des architectures distribuées. Conceptuellement, un agent est un logiciel ou un matériel qui capte de l'information, décide d'agir rationnellement en fonction des données récupérées et déclenche une action pour optimiser ses chances de succès.

Si c'est du matériel, il comprendra des capteurs et des actuateurs. Mais il peut n'être que du logiciel et obtenir des données brutes en entrées et générer des données en sortie.

Un agent réagit donc en fonction de l'environnement et de préférence en temps réel. Les agents intelligents sont intégrés dans des systèmes distribués dénommés systèmes multi-agents avec des agents autonomes, mais reliés et collaborant entre eux.

Les agents sont autonomes, ils appliquent des règles et vont jusqu'à apprendre à les modifier en fonction de l'environnement, ils peuvent être proactifs et pas seulement réactifs à l'environnement, ils communiquent et coopèrent avec d'autres agents et systèmes.

Les solutions d'intelligence artificielle sont ainsi souvent des réseaux multi-agents ! C'est notamment le cas des solutions de RPA (Robotic Process Automation) utilisées notamment dans le back-office des banques et des assurances.

Les réseaux d'agents fonctionnent de manière coordonnée et collective. La coordination de réseaux d'agents est un domaine scientifique à part entière.

On compte notamment les **Distributed Problem Solving** (DPS) qui découpent un problème en sous-problèmes qui sont résolus de manière coopérative entre plusieurs agents reliés les uns aux autres. Ces systèmes sont conçus pour résoudre des problèmes bien spécifiques.

Les réseaux multi-agents servent aussi à faire des simulations, par exemple du comportement de foules dans des gares ou aéroports pour mieux en concevoir les dispositifs de sécurité et d'évacuation. Cela sert aussi dans la simulation du trafic routier et aérien. Ils sont aussi utilisés pour créer des personnages virtuels dans des jeux vidéo. A l'instar des fourmis dans les fourmilières, le comportement collectif de ces agents peut être fascinant. On appelle aussi cela le *swarm computing* ou la *swarm intelligence*. Les SMA (Systèmes Multi-Agents) peuvent aussi servir à simuler l'interaction entre humains et robots pour des applications de cobotique.

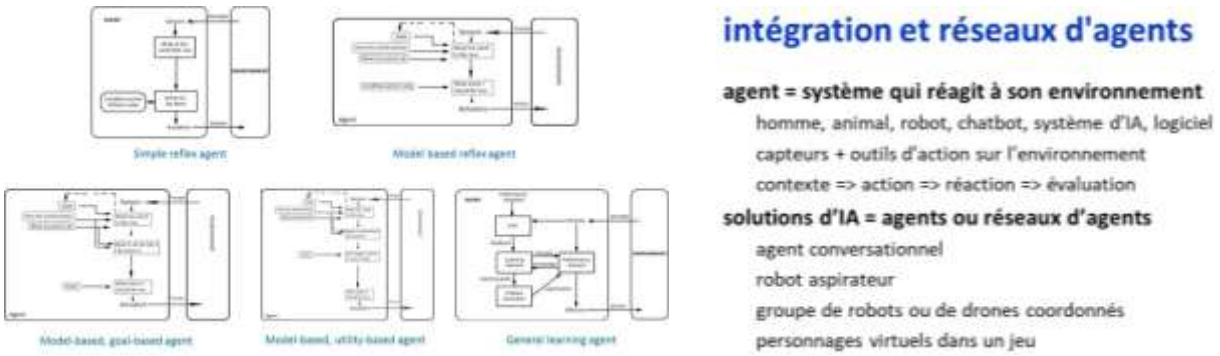
Types d'agents

Les agents sont classifiés par Russell & Norvig dans **Artificial Intelligence – A Modern Approach** (2003-2010) en **types distincts** selon leur niveau d'autonomie et leur mode de prise de décision :

- Les **simple reflex agents** qui comprennent des capteurs, des règles indiquant quelle action mener et des actuateurs pour les déclencher. Ils travaillent en temps réel.
- Les **model based reflex agents** qui ajoutent un moteur d'état capable de mémoriser dans quel état se trouve l'objet et qui évaluent l'impact des actions pour changer d'état.

[260] Voir AI researchers want to study AI the same way social scientists study humans par Karen Hao, MIT Technology Review, avril 2019.

- Les **goal-based agents** qui prennent leur décision en fonction d'un objectif et déterminent une action pour l'atteindre.
- Les **utility-based agents** qui prennent leur décision en fonction d'un but à atteindre qui est plus général.
- Les **learning agents** qui contiennent une fonction d'auto-apprentissage par interactions avec leur environnement.



Vu de haut, les réseaux d'agents ressemblent aux réseaux de neurones mais leur mode de fonctionnement est différent. Un agent peut très bien être lui-même individuellement construit avec un réseau de neurones pour réaliser une tâche spécifique comme la reconnaissance de la parole ou d'images. Et la liaison entre les agents se fait selon des règles, comme dans un moteur de règles ou de workflow !

Les grandes références dans la recherche comprennent notamment **Dinesh Manocha**[261] ainsi que **Serge Hoogendorn** (TU Delft), ce dernier étant spécialisé dans la simulation pour la mobilité urbaine ainsi que **Jacques Ferber** (LIRMM, Montpellier), **Daniel Zucker** (Australie) et **Alexis Drogoul** (IRD, Institut de Recherche et Développement au Vietnam)[262].

On trouve aussi d'excellents chercheurs dans ce domaine au LIP6 de Paris VI, notamment **Amal El Fallah Seghrouchni** et **Nicolas Sabouret**. Notons aussi l'existence de **Golaem** (2009, France, $600K), une spin-off d'Inria spécialiste de la simulation de foules en images de synthèse.

L'équipe **UMMISCO** (Unité de Modélisation Mathématique et Informatique des Systèmes Complexes) fait de la recherche sur le swarm computing et la modélisation des systèmes complexes.

En France, **Yves Demaseau** du Laboratoire d'Informatique de Grenoble (CNRS) est un spécialiste des Systèmes **M**ulti-**A**gents. Il a monté un programme de recherche de 10 ans sur le sujet pour créer un système de SMA d'un million d'agents. Il est par ailleurs président de l'**Association Française pour l'Intelligence Artificielle**.

Outils et applications

Quid des outils de développement associés à la création de réseaux d'agents ? **Il y en a plein**, et notamment en open source, tels que **Soar** de l'Université du Michigan qui servent à mener des tâches et raisonnements généraux ou **REPAST**, de l'Université de Chicago, qui permet de créer des simulations d'interactions de millions d'agents[263].

---

[261] Voir Data-driven Crowd Simulation and Crowd Tracking et la vidéo associée.

[262] Références fournies par Laure Bourgois qui a fait de la recherche dans les SMA et a ensuite créé Codataschool.

[263] La conférence de référence sur le sujet des agents semble être l'AAMAS, la dernière édition ayant eu lieu à Montréal en mai 2019. Voir la liste des papiers publiés à cette occasion. En France, nous avons les Journées Francophones sur les Systèmes Multi-Agents qui ont lieu chaque année pendant l'automne est une bonne source d'information. Voir l'édition 2018.

On peut aussi intégrer **Unity Machine Learning Agents** qui servira notamment aux développeurs de jeux vidéo, très friands de réseaux d'agents ainsi que **JADE** (Java Agent DEvelopment framework) qui vient de Telecom Italia et est open source[264].

Dans un réseau multi-agents, on peut avoir une expertise globale distribuée dans l'ensemble des agents du système. Ce sont des spécialistes qui coopérent pour résoudre un problème général comme dans le diagnostic médical ou la conception d'un produit.

On peut sinon distribuer le problème sur des agents ayant des compétences identiques comme pour la surveillance d'un réseau d'énergie avec une supervision répartie sur chacun des nœuds.

La robotique distribuée fait beaucoup appel aux SMA avec des agents concrets qui se déplacent dans un environnement réel, des ensembles de robots qui coopérent pour accomplir une mission ou une robotique cellulaire faite de l'assemblage de robots dont le mouvement est la conséquence de la coordination d'un ensemble d'agents composant le robot.

Les SMA sont utiles pour traiter des problèmes dans lesquels l'information et/ou les actions et/ou prises de décision sont parcellaires et distribuées ou pour lesquelles il est nécessaire d'avoir une coordination, orchestration et une coopération qu'un algorithme centralisé ne pourrait pas fournir.

On peut aussi utiliser cette approche comme une surcouche pour faire coopérer un ensemble de systèmes. C'est alors un assemblage de briques hétérogènes. Mais ces briques peuvent tout aussi bien être homogènes.

Les solutions d'IA sont des Systèmes Multi-Agents

Dans un agent vocal, un agent va utiliser le texte généré par la reconnaissance puis appliquer un processus de reconnaissance sémantique, puis un autre va traiter la question, fouiller dans une base de données ou de connaissance, récupérer des résultats, un autre va formuler une réponse et la renvoyer à l'utilisateur. Idem pour un système de traduction automatique qui va d'abord analyser la parole avec un premier agent, puis réaliser la traduction avec un second, puis utiliser un troisième agent de "text to speech" pour transformer le résultat de manière audible.

Un robot conversationnel est aussi un réseau d'agents, surtout si on interagit avec la voix avec lui. Les agents sont notamment utilisés dans les systèmes de call centers. Une start-up française s'était lancée - parmi d'autres - sur ce créneau : **Virtuoz**. Elle a été acquise en 2013 par l'américain **Nuance**. Il existait même un concours du meilleur agent de service client en ligne, lancé en 2016 en France avec une trentaine de candidats mais qui n'a visiblement pas été reconduit depuis.

Un robot autonome est aussi un condensé de nombreux agents qui gèrent différents niveaux d'abstraction avec de nombreux capteurs, de la mécanique, des systèmes permettant au robot de savoir où il est, avec quoi il interagit, et qui a des missions à accomplir (aider une personne, conduire un véhicule, etc).

Un robot est particulièrement complexe à mettre au point car il cumule des défis au niveau des capteurs, de l'intégration de ses sens, de la mécanique pour se mouvoir, de la batterie pour son autonomie, et dans l'intelligence artificielle pour piloter l'ensemble et éventuellement interagir à la fois mécaniquement, visuellement et oralement avec son environnement, notamment s'il s'agit de personnes.

[264] Voir aussi le cours Modélisation et Simulation Multi-Agents - Cours 1 - Prolégomènes de Jean-Daniel Kant (103 slides) qui décrit bien la notion de SMA ainsi que Outils et langages de programmation d'applications multi-agents 2017 (46 slides) qui fait un inventaire actualisé des outils de développement de systèmes multi-agents.

C'est dans le domaine de l'**intelligence artificielle intégrative** que des progrès significatifs peuvent être réalisés. Elle consiste à associer différentes méthodes et techniques pour résoudre des problèmes complexes voire même résoudre des problèmes génériques. On la retrouve mise en œuvre dans les agents conversationnels tels que ceux que permet de créer IBM Watson ou ses concurrents.

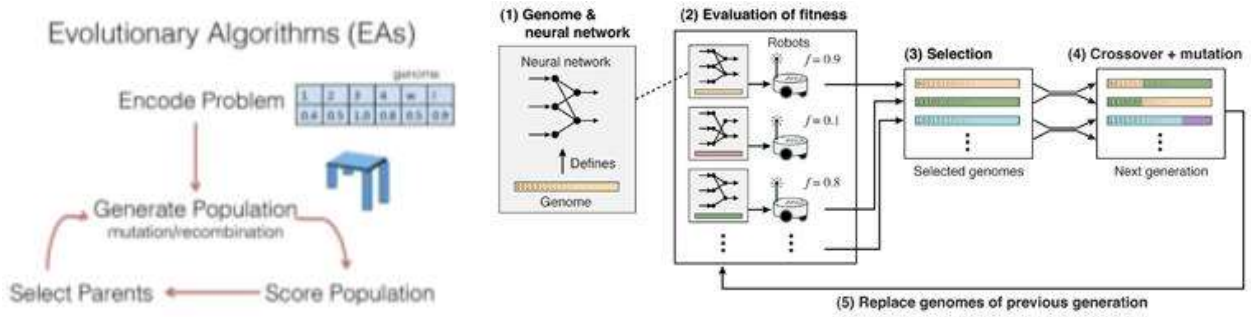
Dans le jargon de l'innovation, on appelle cela de l'innovation par l'intégration. C'est d'ailleurs la forme la plus courante d'innovation et l'IA ne devrait pas y échapper. Cette innovation par l'intégration est d'autant plus pertinente que les solutions d'IA relèvent encore souvent de l'artisanat et nécessitent beaucoup d'expérimentation et d'ajustements.

Cette intégration est un savoir nouveau à forte valeur ajoutée, au-delà de l'intégration traditionnelle de logiciels via des APIs classiques. Cette intelligence artificielle intégrative est à l'œuvre dans un grand nombre de startups du secteur et en particulier dans celles de la robotique.

Le mélange des genres n'est pas évident à décrypter pour le profane : machine learning, deep learning, support vector machines, modèles de Markov, réseaux bayésiens, réseaux neuronaux, méthodes d'apprentissage supervisées ou non supervisées, etc. D'où une discipline qui est difficile à benchmarker d'un point de vue strictement technique et d'égal à égal. Ce d'autant plus que le marché étant très fragmenté, il y a peu de points de comparaison possibles entre solutions. Soit il s'agit de produits finis du grand public comme la reconnaissance d'images ou vocale, et d'agents conversationnels très à la mode en ce moment, soit il s'agit de solutions d'entreprises exploitant des jeux de données non publics.

La programmation génétique

La **vie artificielle** et la **programmation génétique** sont d'autres pans de recherche important connexe aux recherches sur l'IA. Il s'agit de créer des modèles permettant de simuler la vie avec un niveau d'abstraction plus ou moins élevé. On peut ainsi simuler des comportements complexes intégrant des systèmes qui s'auto-organisent, s'auto-réparent, s'auto-répliquent et évoluent d'eux-mêmes en fonction de contraintes environnementales. Et les éléments les moins efficaces de ces systèmes sont éliminés, comme dans le processus de sélection naturelle décrit par Darwin.



Ces systèmes exploitent des algorithmes évolutifs qui sont à la croisée des chemins du deep learning et des réseaux d'agents. Ils consistent à tester différentes combinaisons de réseaux de neurones voire de réseaux d'agents les intégrant pour comparer leur efficacité et conserver les variantes les plus efficaces.

C'est une reproduction informatique du principe de la sélection darwinienne. Reste à s'assurer qu'ils sont efficaces, ce qui est loin d'être évident vu la combinatoire de scénarios qu'ils peuvent être amenés à simuler !

Artificial General Intelligence

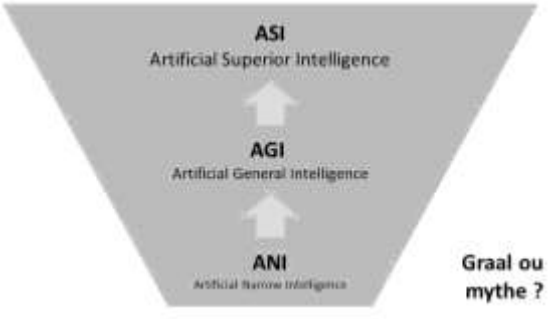
L'AGI nous ramène aux objectifs initiaux des pères de l'IA : intégrer l'intelligence humaine dans des machines, en pièces détachées ou de manière intégrée pour la rendre égale à celle de l'Homme dans ses différentes dimensions, notamment en termes de capacité de raisonnement et d'adaptation.

Cela reste un Graal très lointain voire impossible à atteindre mais divers chercheurs y travaillent. Les approches choisies sont variées avec le rapprochement de l'IA déductive et de l'IA inductive (ou symbolique vs connexionniste)[265], avec la poursuite des progrès dans les neurosciences pour mieux comprendre le fonctionnement du cerveau et enfin, avec des approches complémentaires visant carrément à connecter le cerveau aux IA si ce n'est à le copier intégralement pour le faire tourner dans une machine. Un fantasme plus qu'autre chose compte-tenu de l'infinie complexité du cerveau, de son lien avec le reste du corps humain biologique et avec le monde extérieur.

Trois niveaux de l'IA

Au plus haut niveau conceptuel, on segmente habituellement l'IA en **IA forte** qui imiterait l'intelligence humaine avec capacité de raisonnement généraliste et **IA faible**, qui évoluerait de manière incrémentale à partir d'outils plus élémentaires, soit l'état de l'art actuel.

La distinction entre IA forte et IA faible se retrouve dans cette **classification** de la portée de l'IA avec trois niveaux d'IA : l'ANI, l'AGI et l'ASI, l'ANI correspondant à l'IA faible et les AGI et ASI aux IA fortes. C'est une classification très simpliste et bien trop anthropomorphique des évolutions futures de l'IA, très étroitement liées aux mythes de la **singularité**, qui anticipe ce moment fatidique où une machine atteindra puis dépassera l'intelligence humaine, ce qui n'a aucun sens quand on regarde les choses de près[266].



L'**Artificial Narrow Intelligence** (ANI) correspond à l'état de l'art actuel avec une IA capable de traiter des problèmes des domaines très précis. C'est ce que font aujourd'hui les IA exploitant aussi bien le machine learning que le deep learning ou les moteurs de règles. On peut y mettre en vrac les moteurs de recherche courants, la détection de fraudes bancaires, le credit rating de particuliers, la conduite automatique ou assistée, l'interprétation d'imagerie médicale, Apple SIRI, Amazon Alexa, Microsoft Cortana et Google Assistant.

Si l'IA n'imite pour l'instant pas encore toutes les composantes de l'intelligence humaine, la force brute et l'usage d'éléments techniques dont l'homme ne dispose pas comme la vitesse de traitement et le stockage de gros volumes de données permettent déjà à la machine de dépasser l'homme dans tout un tas de domaines ! Et ce n'est pas nouveau ! Un tableur est déjà des millions de fois plus puissant qu'un humain doté des meilleures capacités de calcul mental ! La mémoire brute d'un Homme est très limitée.

On peut estimer que la mémoire verbale d'une personne moyenne est située aux alentours d'un Go de données ! Mémoire à laquelle il faudrait évidemment ajouter toute la mémoire sensorielle : visuelle, auditive et olfactive, qui est probablement très dense et maillée. Nos souvenirs sont en effet généralement multisensoriels et associatifs, une fonction qui est pour l'instant très difficile à reproduire dans des machines.

---

[265] Voir If AI's So Smart, Why Can't It Grasp Cause and Effect? par Will Knight, mars 2020. Et aussi Principes d'hybridation IA sur une machine parallèle par Aymeric Poulain Maubant, 2020 qui évoque l'histoire de cette hybridation qui commence en 1995.

[266] Si Ray Kurzweil est connu comme le grand promoteur de la singularité, son créateur est en fait Vernor Vinge dans The Coming Technological Singularity: How to Survive in the Post-Human Era en 1993 (12 pages). Il prévoyait l'échéance de la singularité pour 2023. Cette échéance correspondait au passage à la fin de l'humanité. Mais si l'Humanité court à sa perte, ce n'est pas forcément par ce biais là mais bien plus à cause de questions environnementales. Voir Le mythe de la Singularité technologique par Jean-Gabriel Ganascia, septembre 2020.

Dans le second étage, l'**Artificial General Intelligence** (AGI) correspondrait conceptuellement à un niveau d'intelligence équivalent à celui de l'Homme, avec un côté polyvalent, avec la capacité à raisonner, à modéliser les connaissances, analyser des données et résoudre des problèmes variés. L'AGI est en fait dans la continuité des travaux des pionniers de l'IA qui cherchaient à créer des systèmes d'IA capables de résoudre de manière générique toutes sortes de problèmes et en avaient même relancé l'idée au milieu des années 2000 dans l'initiative **HLAI** (Human Level AI).

On peut intégrer dans ce niveau un grand nombre des capacités humaines : l'usage du langage à la fois comme émetteur et récepteur, l'apprentissage par la lecture ou l'expérience, la mémoire et en particulier la mémoire associative, l'usage de la vue et les autres sens, le jugement et la prise de décisions, la résolution de problèmes multi-facettes, la création en général, notamment au niveau conceptuel mais aussi pour résoudre des problèmes physiques, la perception du monde et de soi-même, la capacité à réagir à l'imprévu dans un environnement complexe physique comme intellectuel ou encore la capacité d'anticipation[267]. Et à plus haut niveau, il faudrait intégrer la conscience, les sentiments, la sagesse et la connaissance de soi.

On pourrait y ajouter la capacité à ressentir des émotions que celles-ci soient personnelles (introspection) ou celle d'autres personnes (l'empathie), avoir des envies et des désirs et aussi savoir gérer ses pulsions et agir avec plus ou moins de rationalité. Ou alors, finalement, une machine n'aurait pas du tout besoin de tout cela pour résoudre des problèmes complexes inaccessibles à l'intelligence humaine. Cette liste est très longue ! Pour l'instant, on est encore très très loin de l'AGI, même si certaines de ces capacités notamment linguistiques et de raisonnement général sont en train de voir le jour.

Jusqu'à présent, les solutions d'IA fonctionnaient à un niveau de raisonnement relativement bas. Il reste à créer des machines capables de gérer le sens commun, une forme d'intelligence génétique capable à la fois de brasser le vaste univers des connaissances – au-delà de nos capacités – et d'y appliquer un raisonnement permettant d'identifier non pas des solutions mais des problèmes à résoudre.

Il reste à apprendre aux solutions d'IA d'avoir envie de faire quelque chose. On ne sait pas non plus aider une solution d'IA à prendre du recul, à changer de mode de raisonnement dynamiquement, à mettre plusieurs informations en contexte, à trouver des patterns de ressemblance entre corpus d'idées d'univers différents permettant de résoudre des problèmes par analogie. On pourrait aussi développer des solutions d'IA capables de créer des théories et de les vérifier ensuite par l'expérimentation.

[267] On continue d'en découvrir tous les jours sur les principes biologiques de base de l'intelligence humaine, comme dans Brain Computation Is Organized via Power-of-Two-Based Permutation Logic publié fin 2016. La capacité d'anticipation correspondrait à la notion de « quotient de conscience » (QC). Voir à ce sujet : Le QI est mort, vive le QC ! par Virginie Rio-Jeanne dans HBR France, juillet 2019. Et aussi Will AI achieve consciousness is a wrong question par Daniel Dennett, février 2019.

Pour ce qui est de l'ajout de ce qui fait de nous des êtres humains, comme la sensation de faim, de peur ou d'envie, d'empathie, de besoin de relations sociales, l'IA ne l'intègre pas. Ce n'est d'ailleurs pas nécessaire pour résoudre des problèmes courants auxquels s'attaquent les solutions à base d'IA.

Comme l'indique si bien **Yuval Noah Harari**, l'auteur du best-seller "Sapiens"[268], "*L'économie a besoin d'intelligence, pas de conscience*" (vidéo) ! On pourrait même ajouter, de rationalité. Laissons donc une partie de notre intelligence voire une intelligence plus développée aux machines et conservons la conscience, les émotions et la créativité !

L'avènement éventuel d'une AGI dépend à la fois des progrès matériels et de notre compréhension toujours en devenir du fonctionnement du cerveau humain qui fait partie du vaste champ de la neurophysiologie, coiffant des domaines allant de la neurobiologie (pour les couches "basses") à la neuropsychologie (pour les couches "hautes").

Le fonctionnement du cerveau apparaît au gré des découvertes comme étant bien plus complexe et riche qu'imaginé. Les neurones seraient capables de stocker des informations analogiques et non pas binaires, ce qui en multiplierait la capacité de stockage de plusieurs ordres de grandeur par rapport à ce que l'on croyait jusqu'à il y a peu de temps.

On sait par contre que le cerveau est à la fois ultra-massivement parallèle avec ses trillions de synapses reliant les neurones entre elles mais très lent, avec une horloge tournant au grand maximum à 100 Hz. C'est aussi un engin très efficace du point de vue énergétique, ne consommant que 20W, par heure, soit l'équivalent d'un laptop équipé d'un processeur Intel Core i7 à pleine puissance. Sa puissance de calcul est estimée à 38 PFLOPS même si cela ne veut pas dire grand-chose vu qu'il s'agit d'un engin analogique.

Cette AGI est à mon sens une vue de l'esprit. Elle considère que l'IA ne se mesure que sur une dimension alors que, nous l'avons vu, elle s'évalue sur au moins deux douzaines de dimensions complémentaires. L'alignement de ces dimensions, la capacité de les rendre toutes supérieures à l'Homme dans la machine est une chimère. Nous aurons sans doute très longtemps des machines supérieures à l'Homme dans de nombreuses dimensions et inférieures dans d'autres, ce qui les rendra complémentaires à l'Homme. Des progrès seront certainement faits pour rendre les machines capables de raisonner par analogie pour résoudre des problèmes complexes mais elles n'intégreront pas toutes les composantes de l'intelligence humaine, notamment dans sa relation avec le monde physique.

Au troisième étage de cette fusée simpliste, l'**Artificial Super Intelligence** (ASI) serait une intelligence largement supérieure à l'Homme. Elle aurait même une intelligence infinie, tirant parti de la loi exponentielle de Moore. Une prévision qui se garde bien de rentrer dans les détails logistiques de l'application de cette loi ! L'échéance d'une telle AGI est située entre 2045 et 2062 selon les prévisions[269].

Ce serait la continuité logique et mécanique de l'étape précédente, liée à la puissance des machines qui se démultiplierait et se distribuerait plus facilement que celle d'un cerveau humain avec ses entrées-sorties et ses capacités de stockages et de traitement limitées et locales. Cette intelligence pourrait disposer de capteurs globaux : sur l'environnement, sur l'activité des gens, leurs déplacements, leurs loisirs, leurs états d'âme. Superintelligence va avec superinformation et super big data !

---

[268] Intervenant en juin 2016 dans la conférence USI organisée par Octo Technology à Paris (vidéo).

[269] En voici un exemple avec L'intelligence artificielle égalera l'intelligence humaine d'ici 2062 qui évoque les prévisions du chercheur australien Toby Walsh, auteur de 2062: The World that AI Made. Prévoir une AGI à 44 ans d'échéance semble évidemment bien trop précis. Reste à lire l'ouvrage, ce que je n'ai pas encore pu faire !

À vrai dire, une AGI serait d'emblée largement supérieure à l'Homme car elle accéderait facilement à tout le savoir humain déjà numérisé[270].

À ce niveau, l'intelligence de la machine dépasserait celle de l'homme dans tous les domaines y compris dans la créativité et même dans l'agilité sociale. Ce point de dépassement est une "singularité". Il est évoqué dans de nombreux ouvrages comme **The Singularity is Near** de Ray Kurzweil[271].

Pour de nombreux prospectivistes, l'ASI apparaîtrait très peu de temps après l'AGI, l'ordinateur faisant preuve d'une capacité à se reproduire lui-même, y compris à l'échelle matérielle. C'est évidemment une vue de l'esprit, tout du moins, lorsque l'on observe la manière dont fonctionnent les data-centers. Si ceux-ci étaient entièrement robotisés et alimentés en serveurs et systèmes de stockage par des camions autonomes eux-mêmes alimentés par des usines entièrement autonomes, pourquoi pas. D'où le besoin de préserver un minimum de contrôle humain dans cette chaîne de valeur.

Dans un essai[272], le philosophe australien **David J. Chalmers** propose de tester d'abord l'ASI dans un environnement entièrement virtuel entièrement déconnecté du monde réel pour tester ses aptitudes. Si cela peut rassurer [273]!

Dans la plupart des prévisions sur l'avènement de l'ASI, il est fait état de la difficulté à la contrôler et qu'elle soit même néfaste pour l'humanité malgré son origine humaine. Elles évoquent une course contre la montre entre startups et grandes entreprises pour être les premiers à créer cette ASI. Voire une course face à l'un des plus gros financeurs de l'IA : la DARPA.

Toutes ces conjectures semblent bien théoriques. Elles partent du principe qu'une ASI contrôlerait sans restriction toutes les ressources humaines. Elles s'appuient aussi sur la possibilité que toutes les sécurités informatiques d'origine humaine pourront être cassées par une ASI. C'est une vision dystopique et anthropomorphique du rôle des machines.

Les perspectives de l'AGI et de l'ASI sont surtout influencées par la science-fiction et ses grands auteurs : **Philippe K. Dick** (Blade Runner, Total Recall), **William Gibson** (Neuromancien, qui a inspiré la série de films **Matrix**), **Arthur C. Clarck** (2001 Odyssée de l'Espace), **Gene Roddenberry** (Star Trek), **Vernon Vinge** (le père spirituel du concept de la singularité dans « The coming technological singularity » en 1993, récupéré ensuite par Ray Kurzweil en 1999 dans « The age of spiritual machines »[274]) et bien entendu **Isaac Asimov** (AI, et ses fameuses trois règles de la robotique).

Les prédictions sur l'avènement de l'AGI sont souvent associées à un usage quelque peu abusif et prospectif de la loi de Moore. Or sa pérennité n'est pas garantie, quand bien même on pourrait mettre au point des ordinateurs quantiques dans les 40 ans qui viennent.

---

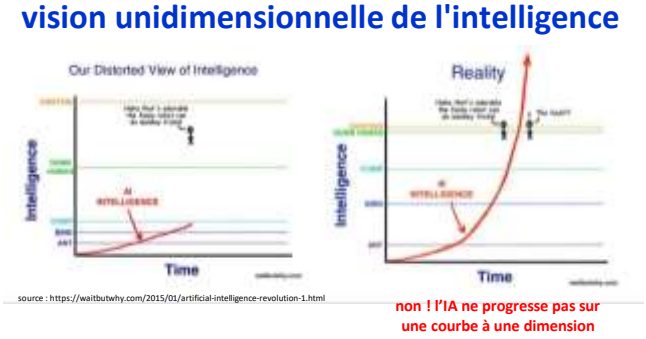
[270] Dans « The inevitable », publié en 2016, Kevin Kelly estime la production de contenu humaine depuis les sumériens à 310 millions de livres, 1,4 milliards d'articles et essais, 180 millions de chansons, 330 000 films de long métrage, 3,5 trillions d'images, un milliard d'heures de vidéo, télévision et courts métrages, et 60 trillions de pages web publiques. Et chaque année, le stock s'agrandirait avec 2 millions de livres, 16 000 films (dont nous ne voyons qu'à peine 1%), 8 millions de chansons et 30 milliards d'articles de blogs. Cela ne comprend pas les données brutes issues d'usages numériques (télécoms, réseaux sociaux, objets connectés). Cela représenterait 50 péta-octets de données. Avec les dernières technologies de stockage SSD, tout cela tiendrait dans un simple rack de data center. Voir le dernier SSD d'Intel. L'intégration de toute cette connaissance dans un réseau de neurones de deep learning se heurterait cependant à des limites techniques pas évidentes à surmonter. Mais avec un bon moteur de recherche, une AGI aurait toutefois une bonne capacité à exploiter cette base de connaissances en fonction des besoins.

[271] La singularité a même sa version russe avec l'initiative 2045 Strategic Social Initiative lancée par l'entrepreneur **Dmitry Itskov** en 2011. L'objet de l'initiative ? Créer les conditions d'une promotion spirituelle, culturelle, éthique, scientifique et technologique de l'humanité. Un nouveau vecteur de civilisation ! Des robots, des mondes virtuels, bref, toutes les nouvelles technologies imaginables du transhumanisme. Voir la vidéo et le site de l'initiative.

[272] Voir The Singularity – A philosophical analysis, David Chalmers (56 pages).

[273] On peut aussi se rassurer avec ce très bon papier Ruper Goodwins paru en décembre 2015 dans Ars Technica UK : Demystifying artificial intelligence: No, the Singularity is not just around the corner.

[274] Voir Technoogical Singularity de Vernon Vinge (16 pages).



Heureusement, l'AGI a aussi eu son camp de détracteurs et dans toutes les décennies récentes (liste non exhaustive *ci-dessous*[275]). En France, le chercheur **Jean-Gabriel Ganacia**[276] auteur du « Mythe de la singularité » est assez remonté contre le mythe de l'AGI.

Tim Dettmers avance que la machine ne pourra pas dépasser le cerveau pendant le siècle en cours. Il démonte les prédictions de Ray Kurzweil[277].

Ragnar Fjelland pense que l'AGI n'est pas réalisable car une IA ne peut avoir une histoire biologique. L'ordinateur n'a pas de compréhension du monde physique car il n'y est pas véritablement intégré et ne sait pas identifier des relations de causalités entre événements. Il revisite la publication controversée d'**Hubert Dreyfus** de 1972, mise à jour plusieurs fois ensuite[278].

Mais il n'est pas nécessaire de maîtriser le niveau d'abstraction le plus bas du cerveau pour en simuler les niveaux élevés, sans passer par un clonage. Comme il n'est pas nécessaire de maîtriser les bosons de Higgs pour faire de la chimie ou comprendre la manière dont l'ADN sert à fabriquer des protéines au sein des cellules !

John Randolph Lucas, Minds, Machines and Gödel, Philosophy XXXVI, 1961.

Hubert Dreyfus, "What Computers Can't Do", 1972.

Roger Penrose, "The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics", 1989.

Nills Nilsson, "Human-Level Artificial Intelligence? Be Serious!", 2005.

David Chalmers, Contemporary Philosophy of Mind: An Annotated Bibliography, Part 4: Philosophy of Artificial Intelligence.

En tout cas, quoi qu'il arrive, l'intelligence d'une machine hyper-intelligente n'aura pas une intelligence similaire à celle de l'homme. Elle sera probablement plus froide, plus rationnelle, moins émotionnelle et plus globale dans sa portée et sa compréhension du monde.

---

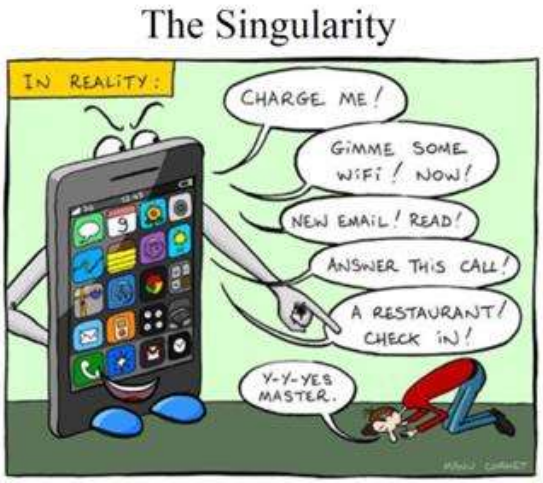
[275] Voir aussi The Emperor of Strong AI Has No Clothes Limits to Artificial Intelligence de Adriana Braga et Robert Logan, 2017 (21 pages).

[276] Voir La Singularité, ça ne tient pas la route ! de Hubert Guillaud, juin 2017.

[277] Dans The Brain vs Deep Learning Part I: Computational Complexity — Or Why the Singularity Is Nowhere Near, 2015. J'avais moi-même émis des doutes sur les exponentielles qui sont la primitive des raisonnements de Kurzweil en avril 2015 dans trois articles sur La dérive des exponentielles. Voir aussi The Seven Deadly Sins of Predicting the Future of AI de Rodney Brooks, septembre 2017.

[278] Voir Why general artificial intelligence will not be realized par Ragnar Fjelland, 2020 (9 pages) et What Computers Can't Do par Hubert Dreyfus, 1972.

L'intelligence artificielle sera supérieure à celle de l'homme dans de nombreux domaines et pas dans d'autres, comme aujourd'hui. Elle sera simplement différente et complémentaire. Tout du moins, à une échéance raisonnable de quelques décennies. Enfin, il reste l'humour. Pour réussir le test de Turing, il existe une solution très simple : rendre les gens moins intelligents ! C'est d'ailleurs l'impact qu'ont souvent les outils numériques, avec les différentes formes d'addiction qu'ils génèrent. Mais pas qu'eux. Le type de médias et contenus consommés ont aussi une forte influence. C'est la thèse ironique de **Piero Scaruffi**[279] et aussi celle de **Nicholas Carr**[280].

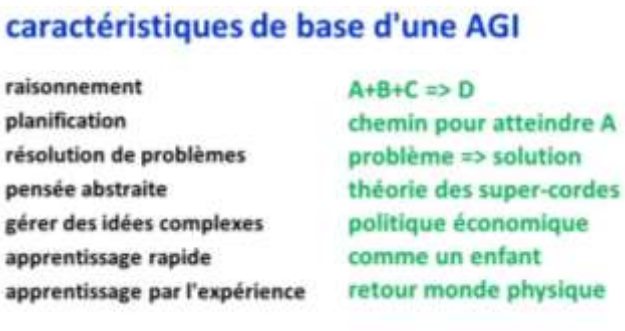


Pourquoi faire ?

Les futurologues qui s'avancent sur l'échéance de l'apparition d'une AGI oublient systématiquement de traiter une question clé : pourquoi faire ? Repartir des besoins et des problèmes permettrait peut-être de circonscrire le sujet et les recherches ! On aimerait bien qu'ils mettent sur la table quelques problèmes clés permettant d'améliorer le sort de la planète et de ses habitants. A la place, on a droit à des prévisions anxiogènes sur des AGI qui emprunteraient à l'Homme tout ce qu'il a de mauvais.

Quelques cas pratiques d'usage de l'AGI ont été définis par le passé mais ils sont très généraux, comme le **test de Turing** (agent conversationnel textuel que l'on ne peut pas distinguer d'un humain), le **test de la machine à café** de Steve Wozniak (un robot peut entrer dans un logement, trouver la machine à café, l'eau, le café et la tasse, faire le café et le servir), le **test de l'étudiant** robot (capable de suivre de cours et de passer avec succès les examens), celui du **chercheur** (capable de mener des travaux de recherche, de soumettre une thèse et d'obtenir son doctorat) et enfin, celui du **salarié** de tel ou tel métier, peut-être le cas le plus facile pour certains métiers répétitifs.

Les choses pourraient se compliquer pour une AGI si on lui demandait de démontrer le théorème d'incomplétude de **Godel** selon lequel "*dans n'importe quelle théorie récursivement axiomatisable, cohérente et capable de « formaliser l'arithmétique, on peut construire un énoncé arithmétique qui ne peut être ni prouvé ni réfuté dans cette théorie"* ou encore du dernier théorème de **Fermat** ($x^n + y^n = z^n$ est impossible pour un entier n supérieur à 2 ; x, y et z étant également des nombres entiers).

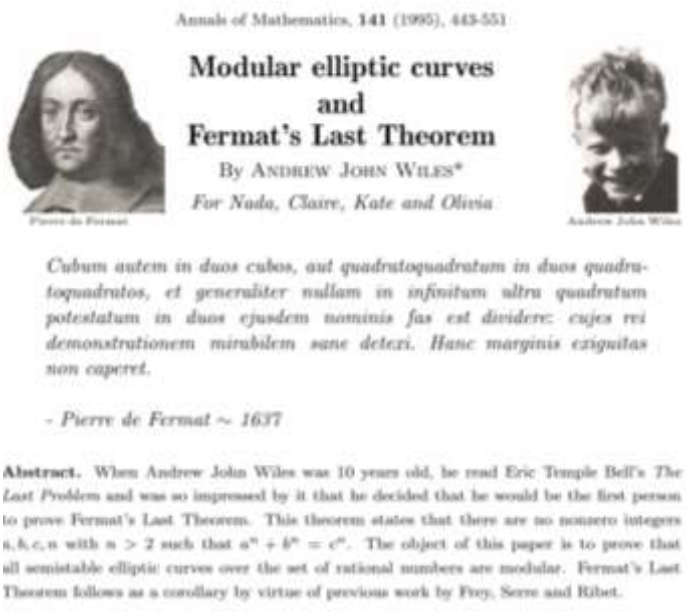


[279] Dans Artificial Intelligence and the Singularity, octobre 2014.

[280] Dans Is Google making us stupid, dans The Atlantic, juillet 2008.

La conjecture de Fermat, devenue théorème, est fascinante. Ce dernier a été démontré par l'Anglais **Andrew John Wiles** et après des années d'efforts de plusieurs mathématiciens. Sa démonstration publiée dans les annales de mathématiques en 1995 fait 109 pages et fait appel à de nombreux concepts incompréhensibles au commun des mortels, y compris pour votre serviteur passé par les classes préparatoires scientifiques au 20e siècle qui s'y perd à la troisième ligne de la démonstration[281].

Un défi a été lancé en 2005 par un certain Jan Bergstra pour démontrer le théorème de Fermat avec un ordinateur et il reste toujours à relever[282]. À vous de jouer si cela vous tente ! Le jour où une IA démontrera le théorème de Fermat toute seule, il y aura vraiment avoir de quoi être bluffé !



Annals of Mathematics, **141** (1995), 443-551

Modular elliptic curves and Fermat's Last Theorem

By Andrew John Wiles*

For Nada, Claire, Kate and Olivia

Cubum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos, et generaliter nullam in infinitum ultra quadratum potestatum in duos ejusdem nominis fas est dividere: cujus rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

- Pierre de Fermat ~ 1637

Abstract. When Andrew John Wiles was 10 years old, he read Eric Temple Bell's *The Last Problem* and was so impressed by it that he decided that he would be the first person to prove Fermat's Last Theorem. This theorem states that there are no nonzero integers a, b, c, n with $n > 2$ such that $a^n + b^n = c^n$. The object of this paper is to prove that all semistable elliptic curves over the set of rational numbers are modular. Fermat's Last Theorem follows as a corollary by virtue of previous work by Frey, Serre and Ribet.

Autre défi à relever pour une intelligence artificielle, résoudre les six défis mathématiques restants du **Clay Institute** lancés en 2000[283], avec $1M de récompense à la clé pour chacun d'entre eux (lien). Cela comprend la démonstration des équations de Navier-Stokes de la mécanique des fluides, la démonstration de P=NP ou P<>NP dans la théorie de la complexité[284] ou la démonstration de la conjecture de Poincaré !

Tant que l'on y est, on pourrait ajouter l'unification des théories de la physique des particules ou résoudre l'hypothèse de Riemann qui fait partie des 23 problèmes de Hilbert énoncés en 1900 et qui n'est toujours pas résolue. Je pose la question car il resterait à définir les contours d'une AGI capable de démontrer ces théorèmes ou ces conjectures.

Des chercheurs de Google Research ont commencé en 2017 à s'attaquer à ce problème en créant une base de connaissances ouverte classifiée exploitant les démonstrations de 11 400 théorèmes et plus de deux millions d'étapes intermédiaires de démonstrations[285]. Reste à l'exploiter pour démontrer d'autres théorèmes et conjectures (théorèmes non encore démontrés) !

Bref, lorsque l'on parle d'AGI ou d'ASI, il vaut mieux non pas se comparer à l'Homme moyen, mais aux scientifiques et inventeurs les plus chevronnés de l'Histoire !

Où en est-on aujourd'hui ? En 2015, une IA atteignait le QI d'un enfant de 4 ans dans un système de questions/réponses[286]. En 2019, l'IA **Aristo** d'AI2 (Allen Institute for Artificial Intelligence) réussissait à passer un test de science de classe de 4ième comprenant 119 questions avec 90% de bonnes réponses. Sur un test de classe de terminale, le taux de réussite était de 83%.

---

[281] Voir la démonstration : Modular elliptic curves and Fermat's Last Theorem, Andrew John Wiles, 1995 (109 pages). Une première demonstration du théorème avait été publiée en 1993, mais contenait une erreur qui a nécessité deux ans de travail pour être corrigée.

[282] Voir Computer verification of Wiles' proof of Fermat's Last Theorem. Depuis, une IA de Google AI était capable de démontrer 1200 théorèmes mathématiques plus simples. Voir Google AI system proves over 1200 mathematical theorems « Math Scholar par David H Bailey, avril 2019.

[283] Voir Clay Mathematics Institute – Millenium Problems. Un seul de ces défis à été relevé pendant les 18 ans qui nous séparent du lancement de cette initiative.

[284] J'évoque la question dans Comprendre l'informatique quantique – Complexité, juillet 2018.

[285] Voir HolStep: a Machine Learning Dataset for Higher-Order Logic Theorem Proving, 2017, Cezary Kaliszyk de l'Université d'Innsbruck, François Chollet et Christian Szegedy.

[286] Voir IQ Test Result: Advanced AI Machine Matches Four-Year-Old Child's Score, octobre 2015, qui fait référence à Measuring an Artificial Intelligence System's Performance on a Verbal IQ Test for Young Children, par Stellan Ohlsson & Al, 2015 (17 pages).

Mais ces tests n'étaient pas complets et ne comprenaient pas ceux portant sur l'interprétation de diagrammes et sur des réponses à des questions ouvertes. Aristo s'appuie sur huit différents agents construits sur ELMo d'AI2[287] et sur **BERT** de Google[288] : un agent qui recherche des réponses dans une base de données comme dans un chatbot de premier niveau, un agent qui vérifie l'association de concepts et un autre qui gère un raisonnement qualitatif. Ces solutions restent pour l'instant des succédanés d'intelligence humaine qui manquent de sens commun pour réellement résoudre une grande variété de problèmes[289].

L'autre attente potentielle vis à vis de l'IA est d'en faire un outil créatif. Tout comme l'intelligence, la notion de créativité et d'innovation est largement débattue. Elle est décrite prosaïquement comme un processus de recherche et de recombinaison d'éléments existants, une méthode mise en œuvre dans les réseaux de neurones adversariaux (GAN). L'innovation est un processus de résolution de problèmes existants ou latents. Mais les IA d'aujourd'hui ne savent pas encore définir des problèmes et leur importance, ce d'autant plus qu'ils ne sont forcément détectables qualitativement dans les données du passé. Une IA isolée est rarement créative. En effet, le processus de créativité nécessite un aller et retour avec le monde réel, à la fois pour s'en inspirer et aussi pour obtenir son retour. C'est via cette confrontation qu'émergent l'adoption, les effets de mode et l'appréciation du public. Par certains côtés, c'est de l'apprentissage par renforcement piloté par les Humains. Finalement, pour encore longtemps, l'IA sera l'ingrédient de boîtes à outils exploitées par les créatifs, comme les pinceaux et la plume[290].

Autre exemple de chimère avec l'IA : son éventuelle capacité à prévoir le futur. Un journaliste avait tenté l'exercice avec GPT-2 fin 2019 pour prédire les événements de 2020, autour de l'IA et en général.

En février, prédire l'année 2020 ne pouvait pas s'appuyer sur du machine learning, quelle que soit la méthode utilisée. Une IA ne prédit rien du tout en pratique. Les réponses étaient toutes faites à partir de textes d'origine humaine. Par contre, elle avait vu juste sur un point : la défaite de Donald Trump lors de l'élection présidentielle de 2020. À l'époque, une majorité de commentateurs le donnaient gagnant. Reste à saisir les éléments qui lui ont permis d'affirmer cela. Bien entendu, cette IA n'avait pas prévu l'impact mondial de l'épidémie du covid-19[291].

Comprendre le cerveau

Le concept même d'IA ne fait pas l'unanimité dans sa définition. Pour les puristes, un simple réseau de neurones ou un système de reconnaissance d'images ne relève pas à proprement parler de l'IA. Tout dépend de la définition que l'on se donne de l'IA, et notamment si la définition est anthropocentrée ou pas. C'est un peu comme la magie. Tant que l'on ne connaît pas le truc, c'est de la magie voire de l'art. Une fois qu'on le connaît, c'est une technique, souvent très simple, si ce n'est évidente.

---

[287] Voir https://allennlp.org/elmo.

[288] Voir Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing, novembre 2018.

[289] Voir AI Can Pass Standardized Tests—But It Would Fail Preschool par Melanie Mitchell, septembre 2019. Voir aussi Two rival AI approaches combine to let machines learn about the world like a child par Will Knight, avril 2019, qui fait référence à The Neuro-Symbolic Concept Learning : Interpreting Scenes, Words and Sentences from Natural Supervision, 2019 (28 pages).

[290] Voir Can AI ever rival human creativity? Here's what the science says par Tim Schweisfurth et René Chester Goduscheit, juin 2020.

[291] Voir An artificial intelligence predicts the future - What would an artificial intelligence think about the year ahead? We decided to ask one dans The Economist, février 2020 et This Is What an AI Said When Asked to Predict the Year Ahead par Vanessa Bates Ramirez, février 2020. Le journaliste qui a réalisé cette interview explique dans How I (sort of) interviewed an artificial intelligence, par Tom Standage, décembre 2019, comment il s'y est pris. Pour chaque question, GPT-2 renvoyait cinq réponses et il choisissait celle qui lui semblait la plus appropriée et sans l'éditer. Il ajoutait un biais humain de tri.

L'intelligence humaine est un peu du même ressort quand on n'en connaît pas le fonctionnement exact. Elle préserve ce côté mystérieux et inimitable, presque immatériel, comme une âme qui n'aurait pas d'existence physique. L'intelligence humaine est magique tant qu'on a du mal à en expliquer le fonctionnement ou à connaître son origine exacte. C'est à la fois une affaire de neurosciences et de biologie (à bas niveau) et de sciences humaines (à haut niveau)[292].

La notion même d'intelligence humaine est un sujet débattu à l'infini. Elle comprend souvent une notion d'adaptation et d'apprentissage. D'autres la définissent comme le moyen de maximiser sa future liberté d'action[293].

Au gré des découvertes en neurobiologie et en sciences cognitives, cette magie perd petit à petit de son lustre. L'homme n'est après tout qu'une machine biologique très sophistiquée issue de l'évolution. Certes, une machine complexe, une machine dont le fonctionnement dépend d'un très grand nombre de paramètres environnementaux et de l'accumulation d'expériences, mais une machine tout de même.

C'est la première d'entre elles qui soit d'ailleurs capable d'en comprendre son fonctionnement interne ! Une plante ne connaît rien aux principes de la photosynthèse ou au métabolisme des glucides dans les cellules vivantes décrit par le cycle de Krebs.

Doit-on absolument chercher à copier ou imiter le cerveau humain pour créer des solutions numériques ? Dans quel cas l'imitation est-elle utile et dans quels cas l'inspiration seulement nécessaire ? Doit-on chercher à créer des machines plus intelligentes que l'homme dans *toutes* ses dimensions ?

L'exemple de l'aviation couramment utilisée par Yann Le Cun peut servir de bonne base de réflexion. L'avion s'inspire de l'oiseau mais ne l'imite pas pour autant. Les points communs sont d'avoir des ailes et d'utiliser la vitesse et la portance des ailes pour voler.

Le concept diverge alors rapidement : les avions n'ont pas d'ailes mobiles faites de plumes ! En lieu et place, leurs ailes sont généralement fixes et les moteurs sont à hélice ou réacteurs. L'avion dépasse largement l'oiseau dans la vitesse (supersonique pour les avions militaires), la taille (B747, A380, Galaxy C5, Antonov 124), la capacité d'emport (qui se mesure en dizaines de tonnes), l'altitude (10 km pour un avion de ligne) et la résistance du froid (il y fait environ -50°C, ce qu'un organisme biologique développé peut difficilement supporter longtemps, même avec un bon plumage). Les avions sont par contre très inférieurs aux oiseaux côté efficacité énergétique et flexibilité, même si la densité énergétique de la graisse animale est voisine de celle du kérosène (37 vs 43 Méga Joules/Kg).

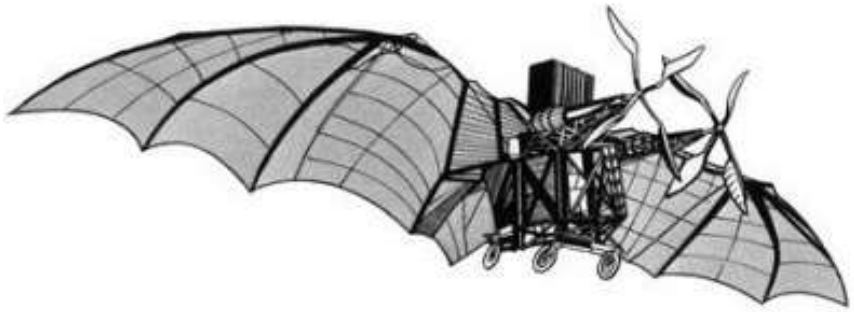
Le bio-mimétisme a été utile au début pour conceptualiser l'avion, que ce soit dans les schémas de Léonard de Vinci ou de l'avion de Clément Ader qui étaient très proches de l'oiseau. Si la motorisation d'un avion est très différente de celle des oiseaux qui battent de l'aile, les plumes se déployant au moment de l'atterrissage et du décollage sont cependant réapparues sous la forme des volets hypersustentateurs. Ils ont été inventés par Boeing pour les 707 lancés à la fin des les années 1950 (**description**) et dont la forme la plus élaborée a été intégrée aux Boeing 747 (*ci-dessous à droite*), dont les premiers vols ont eu lieu en 1969.

L'aigle est l'un des oiseaux les plus rapides au monde, atteignant 120 km/h, le faucon pélerin pouvant même faire du 389 km/h en piqué. Un avion de ligne classique atteint 1000 Km/h et il touche le sol, volets hypersustentateurs déployés, à environ 200 km/h. Un A380 décolle en 2700 m et atterrit sur 1500 m. Un aigle se pose en quelques secondes et presque n'importe où ! C'est la puissance et la capacité d'emport contre la flexibilité. A contrario, l'avion de Clément Ader (*ci-dessous*, à gauche) qui imitait de trop près la chauve-souris n'a pas connu un grand succès !

---

[292] Voir à ce sujet l'une des références des sciences cognitives, Stanislas Dehaene, avec ses cours au Collège de France. Il y décrit notamment l'origine du langage et des pensées.

[293] Vu dans Can't Define AI? Try Defining Intelligence par Ron Schmelzer, février 2020.



Il faut se pencher du côté des drones de poche pour retrouver une part de la flexibilité des oiseaux mais leur autonomie est généralement bien plus limitée que celles des oiseaux, surtout les oiseaux migrateurs qui peuvent voler plusieurs heures d'affilée avant de se reposer au sol.

L'IA suit un chemin voisin dans le biomimétisme : certaines caractéristiques du cerveau des mammifères sont imitées dans les réseaux de neurones et le deep learning. Mais des différences fondamentales font diverger l'intelligence humaine et celle de la machine : à la fois ses entrées et sorties tout comme la structure de sa mémoire et du raisonnement.

La machine se distingue pour l'instant par la capacité de stockage et d'analyse d'immenses volumes d'informations et par sa puissance de calcul brute. L'homme dispose de capteurs sensoriels en quantité astronomique qu'aucun objet connecté n'égale à ce stade, ce qui, associés au cortex, lui procure une mémoire sensorielle qui accumule les souvenirs pendant toute son existence, provenant des entrées/sorties que sont les nerfs optiques, auditifs, gustatifs et olfactifs, ainsi que ceux qui gèrent le toucher, faits de millions de neurones irrigant en parallèle notre mémoire sensorielle[294]. C'est une force et une faiblesse. Nos émotions liées à cette mémoire sensorielle génèrent la peur de certains risques et des prises de décisions pouvant être irrationnelles. Ensuite, le niveau de complexité du cerveau dépasse l'entendement.

Il n'empêche que, par la force brute, l'IA dépasse déjà l'Homme dans tout un tas de domaines, notamment lorsqu'il faut "cruncher" de gros volumes de données qui nous échappent complètement. Quand elle a accès à de gros volumes de données comme dans l'oncologie ou en exploitant les données issues d'objets connectés, l'IA peut faire des merveilles. Elle est d'ailleurs plutôt inopérante sans données. Elle ne sait pas encore quoi chercher ni prendre d'initiatives. Et les algorithmes sont encore très limités car les données de notre vie ne sont, heureusement, pas encore consolidées. Cela explique les limites de ces algorithmes de recommandation qui ne savent pas ce que j'ai déjà vu ou fait et ne sont pas près de le savoir. Ils ne peuvent donc pas faire de recommandation totalement pertinente. Le jour où toute notre vie sera suivie par des objets connectés depuis la naissance, il en sera peut-être autrement.

Qu'en est-il du raisonnement humain ? Celui-ci ne semble pas hors de portée des machines. On arrive petit à petit à le modéliser pour des tâches très spécialisées. Mais l'IA manque encore de souplesse et de capacité d'adaptation à une grande variété de situations. Bref, de jugeote !

Mais il n'est pas inconcevable d'arriver à fournir une intelligence générique à une machine. On y arrivera par tâtonnements, par intégration de briques algorithmiques et logicielles disparates, et pas seulement via la force brute de la machine[295].

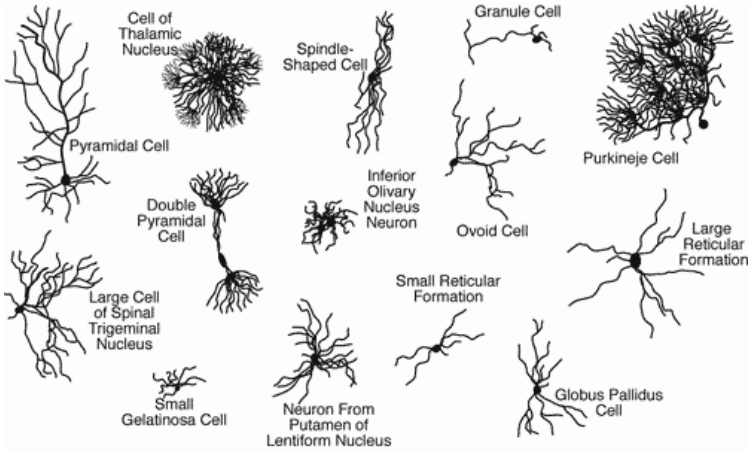
---

[294] Brain Facts and Figures contient un impressionnant dimensionnement de l'ensemble du système nerveux humain. Quelques données clés : la moelle épinière qui relie le cerveau à l'ensemble du corps contient un milliard d'axones. Chaque rétine contient 5 à 6 millions de cones qui captent la couleur et de 120 à 140 millions de bâtonnets qui captent l'intensité de la lumière et peuvent capter un seul photon. Le nerf optique contient 1,2 millions d'axones. Le cortex visual contient 538 millions de neurones. Dans l'oreille, la cochlée contient 15500 cellules cillées. Le nerf auditif contient 30 000 axones. Le cortex auditif contient 100 millions de neurones. Bref, côté sens, le cerveau est compliqué !

[295] Voir cette excellente présentation : Reverse-Engineering the Brain 2018 (192 slides) qui décrit l'état de l'art de la compréhension du fonctionnement du cerveau et essaye de comprendre l'origine de la plasticité cérébrale.

Je vais maintenant m'intéresser au fonctionnement du cerveau pour évaluer sa complexité et la difficulté à en modéliser le comportement dans de l'IA. Celui-ci contient plusieurs centaines de types de neurones différents[296], l'illustration *ci-contre* n'en présentant que quelques grandes variantes.

Le cervelet contient notamment ces étonnantes cellules de Purkineje, avec leur arbre de dendrites reliées avec jusqu'à 200 000 autres neurones, et qui contrôlent les mouvements appris.



source du schéma sur quelques exemples de neurones braux : **http://neuromorpho.org**

Cette complexité se situe aussi au niveau moléculaire avec de nombreuses protéines et hormones intervenant dans la transmission d'influx neuronaux[297]. Parmi les 25 000 gènes de nos cellules, 6000 sont spécifiques au fonctionnement du cerveau et leur expression varie d'un type de neurone à l'autre et en fonction de leur environnement !

Chaque gène entraîne la création d'une grande variété de protéines du fait du principe de l'épissage qui fait se combiner dans des ordres variables les différentes composantes de l'ADN des gènes (les exons). C'est dire la richesse de la soupe de protéines qui gouverne le cerveau, dont l'actine qui structure la forme mouvante des neurones !

Le cerveau d'un fœtus comprendrait plus de mille milliards de neurones, qui meurent rapidement. On perd en fait des neurones dès sa naissance, comme si une matrice s'évidait pour prendre forme progressivement au gré des apprentissages. Le cerveau d'un enfant comprendrait plus de 100 milliards de neurones, et plus de 15 trillions de synapses et 150 milliards de dendrites.

Un cerveau adulte comprend environ 86 milliards de neurones dont 16 milliards dans le cortex et environ 56 milliards dans le cervelet. Ces neurones sont reliés entre eux par environ 600 trillions de synapses (liaisons neurones / neurones via les terminaisons multiples des axones qui sortent de neurones et se connectent aux dendrites proches du noyau d'autres neurones), et 300 milliards de dendrites (les structures des neurones sur lesquelles se trouvent les synapses). Le ventre contient aussi 200 millions de neurones.

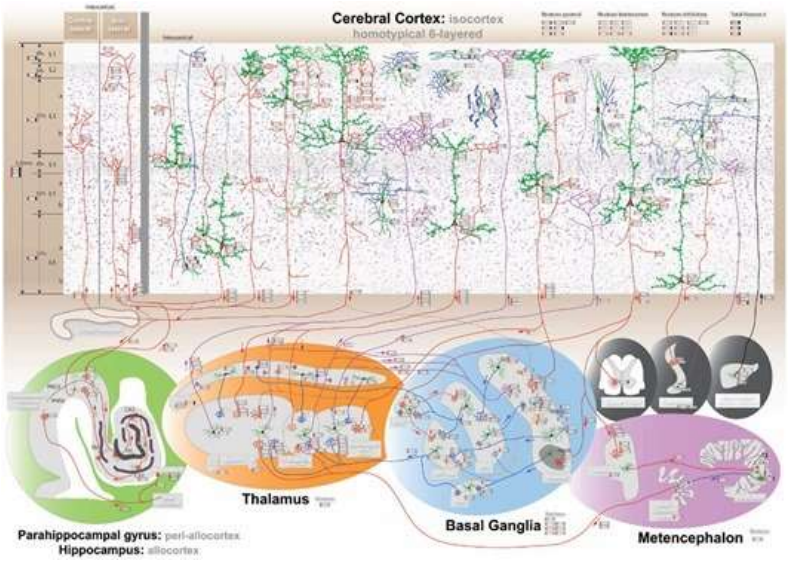
Par comparaison le cerveau d'un chat comprend 760 millions de neurones et 10 trillions de synapses et celui d'un chien comprend 530 millions de neurones.

Les 20 Watts consommés par heure par le cerveau sont fournis sous forme d'hydrates de carbone (glucoses) via la circulation sanguine, ce qui en fait une "machine" très efficace côté consommation énergétique. Dans son développement à partir de la naissance, le cerveau perd des neurones mais gagne des liaisons entre elles, et ce, toute la vie, même si le processus se ralentit avec l'âge, même sans maladies neurodégénératives.

[296] Voir Neurons, Synapses, Action Potentials, and Neurotransmission.

[297] La complexité protéinaire des neurons est décrite dans **Deep Molecular Diversity of Mammalian Synapses: Why It Matters and How to Measure It**.

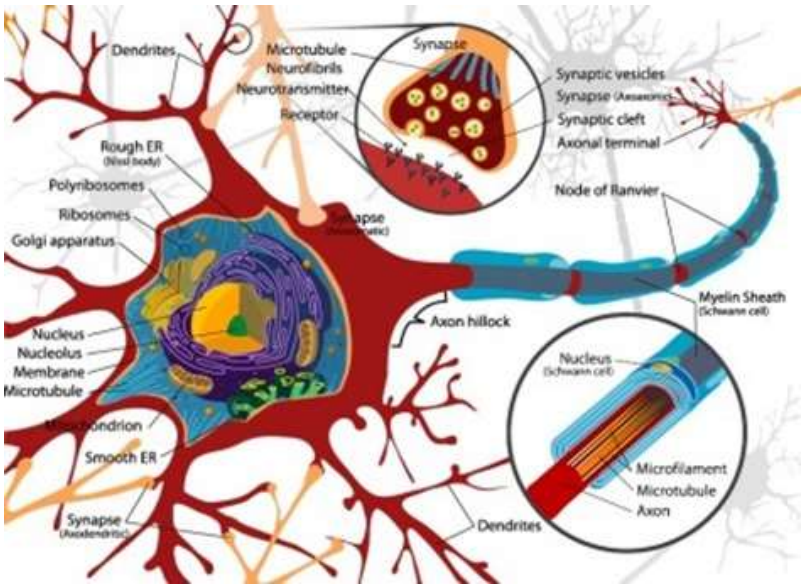
Un neurotransmetteur arrivant via une synapse peut déclencher une cascade de réactions en chaînes dans le neurone cible qui va réguler l'expression de gènes et produire des protéines de régulation qui vont modifier le comportement des dendrites dans la réception des signaux issus des axones. Qui plus est les dendrites – les récepteurs dans les neurones – ont des formes et des comportements variables. Bref, nous avons un système de régulation des plus complexes qui n'a pas du tout été intégré dans les modèles Kurzweiliens !



(source de l'illustration ci-dessus)

Le cervelet cogère avec le cortex les automatismes appris comme la marche, la préhension, les sports, la conduite, le vélo, la danse ou la maîtrise des instruments de musique. Il intègre la perception sensorielle avec la commande motrice provenant du cortex. Un neurone du cervelet contient environ 25 000 synapses le reliant aux terminaisons d'axones d'autres neurones, avec l'exception des cellules de Purkinje qui ont 200 000 connexions.

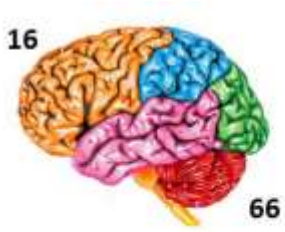
Les neurones du cortex gèrent les sens et l'intelligence. Ils comprennent chacun de 5000 à 15 000 synapses. Le cerveau est aussi rempli de cellules gliales qui alimentent les neurones et en contrôlent le fonctionnement via la myéline qui entoure les axones et divers autres mécanismes de régulation. Il y en a au moins autant que de neurones dans le cerveau, ce qui ajoute un niveau de complexité de plus. Il faut ajouter le rôle de buffer mémoire de l'hippocampe, le vidage de ce buffer s'effectuant pendant le sommeil. Cela explique qu'une bonne qualité et durée de sommeil entretiennent bien sa mémoire.



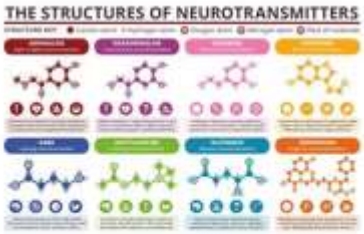
(**source** du schéma qui l'explique très bien)

Enfin, via le système nerveux sympathique et parasympathique, le cerveau est relié au reste des organes, dont le système digestif ainsi qu'à tous les sens et notamment le toucher.

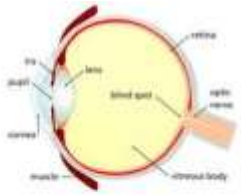
Le cerveau est imbattable dans sa densité, sa compacité et son parallélisme. Par contre, les ordinateurs nous dépassent dans leur capacité de stockage et de traitement de gros volumes de données. Si l'on a toujours, et pour bien longtemps encore, du mal à scanner un cerveau au niveau des neurones, il n'en reste pas moins possible d'en comprendre le fonctionnement par tâtonnements.



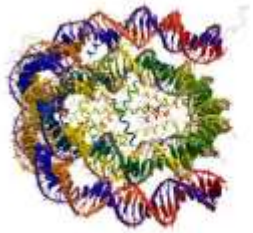
86 milliards de neurones
600 trillions de synapses



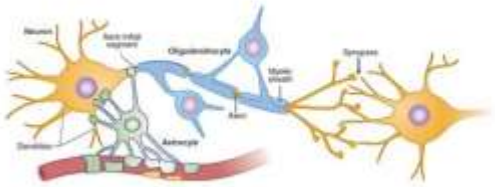
8 neurotransmetteurs différents



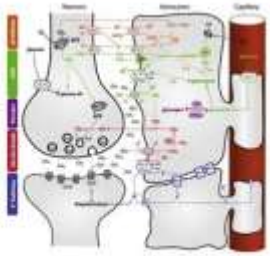
6,5 millions de cônes
90 millions de bâtonnets
1 million d'axones



6000 gênes spécifiques



88 milliards de cellules gliales dans le cerveau



des interactions complexes

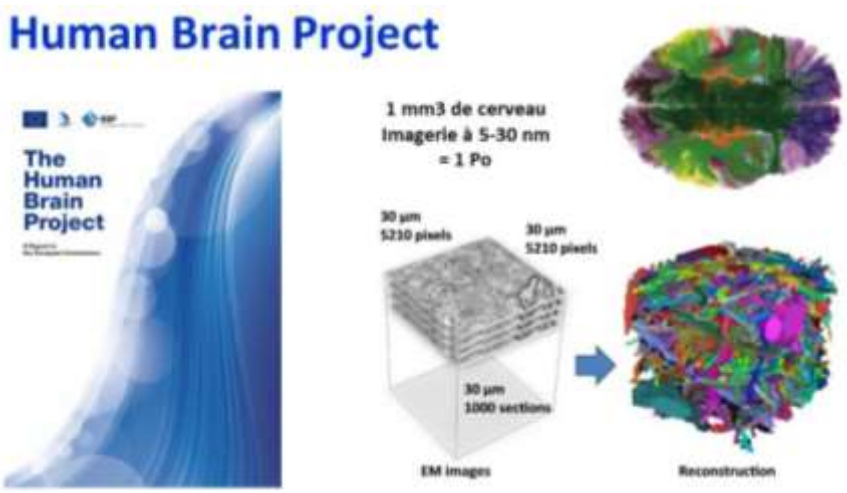
Les neurosciences continuent de progresser régulièrement de ce point de vue-là. On comprend petit à petit comment fonctionnent les différents niveaux d'abstraction dans le cerveau, même si les méthodes scientifiques de vérification associées restent assez empiriques, réalisées le plus souvent sur des souris.

Comprendre le cerveau en modélisant son fonctionnement reste cependant un objectif de nombreux chercheurs. L'idée n'est pas forcément de le copier, mais au moins de mieux connaître son fonctionnement pour découvrir des traitements de certaines pathologies neurodégénératives[298].

De nombreuses initiatives de recherche nationales et internationales ont été lancées dans ce sens. Elles sont issues d'Europe, des USA, mais aussi du Japon, d'Australie, d'Israël, de Corée et d'Inde.

Le projet européen **Human Brain Project** qui est un des trois « flagships » de l'Union Européenne avec les nanotechnologies et l'informatique quantique vise à simuler numériquement le fonctionnement d'un cerveau.

Il a été lancé après la réponse à un appel d'offres par Henry Markram de l'EPFL de Lausanne.



Ce chercheur est à l'origine du **Blue Brain Project** lancé en 2005, qui vise à créer un cerveau synthétique de mammifère.

[298] La tâche est loin d'être évidente. Voir Intelligence artificielle : « La complexité de la modélisation du cerveau humain a toujours été sous-estimée » par Claire Gerardin dans Le Monde, mai 2019.

Disposant d'un budget communautaire de 1Md€ étalé sur cinq ans, le Human Brain Project ambitionnait de manière aussi large que possible d'améliorer la compréhension du fonctionnement du cerveau, avec en ligne de mire le traitement de pathologies neuro-cérébrales et la création d'avancées technologiques dans l'IA.

La cartographie à bas niveau du cerveau humain s'appuie sur le laboratoire **Neurospin** du CEA à Saclay qui est en train de mettre en place le système d'imagerie fonctionnelle par résonnance magnétique nucléaire le plus puissant du monde avec une résolution d'un dixième de millimètre[299]. Inria contribue aussi au projet via son équipe Parietal qui travaille sur le projet Individual Brain Charting de cartographie des fonctions cognitives du cerveau humain. La cartographie est réalisée à partir d'IRM fonctionnelle captées lors de tâches comportementales[300].

A l'aide d'un supercalculateur Blue Gene d'IBM faisant tourner le logiciel de réseau de neurones de Michael Hines, le projet vise ensuite à simuler de manière aussi réaliste que possible des neurones[301] et un cerveau, en partant de celui de petits animaux comme les rats.

Le projet est critiqué ici et là[302]. Il fait penser un peu à Quaero par son aspect disséminé. Les laboratoires français ont récolté 78M€ de financement, notamment au CEA, tandis que ceux d'Allemagne et de la Suisse se sont taillés la part du lion avec respectivement 266M€ et 176M€. On se demande qui fera l'intégration !

C'est plutôt un projet de big data qui s'éloigne du cerveau. En effet, les modèles de simulation ne s'appuient plus du tout sur la connaissance biologique actualisée que l'on a du fonctionnement des neurones dans le cerveau.

Les USA ne sont pas en reste avec la **BRAIN Initiative** annoncée par Barack Obama en 2013 et pilotée par le NIH (National Health Institute, équivalent US de l'INSERM français). Elle visait à mieux comprendre le fonctionnement du cerveau. L'objectif annoncé semble plus opérationnel que celui des européens : mieux comprendre les maladies d'Alzheimer et de Parkinson ainsi que divers troubles neuronaux. Le budget annuel était de l'ordre de $100M, donc, in fine, du même ordre de grandeur que celui du Human Brain Projet. Parmi les projets, on trouve des initiatives en nanotechnologies pour mesurer l'activité individuelle de cellules nerveuses, à commencer par celles des mouches drosophiles. Le plan initial est arrivé à son terme en 2018 et le NIH a lancé un appel à propositions pour le renouveler sur 5 ans.

Dans le cadre de la BRAIN Initiative, la **DARPA** a démarré en 2016 un projet visant à lire simultanément l'état d'un million de neurones d'ici 2020 dans ce que l'on appelle la BCI (Brain Computer Interface)[303].

On peut aussi citer le **Human Connectome Project**, lancé en 2009, un autre projet américain, financé par le NIH comme la BRAIN Initiative, et qui vise à cartographier avec précision les différentes régions du cerveau.

---

[299] L'équipe de Neurospin a publié en août 2018 Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping, décrit dans Individual Brain Charting : une cartographie cérébrale à haute résolution des fonctions cognitives. L'état de l'art actuel est une cartographie du cerveau en IRM fonctionnelle avec une résolution spatiale de 1,5 mm.

[300] Voir L'imagerie haute résolution cartographie le cerveau humain pour créer un atlas des fonctions cognitives cérébrales, par Elisa Riva, Inria, juin 2019.

[301] J'ai eu l'occasion de rentrer un peu plus en détails dans ce projet dans Ces startups qui veulent bidouiller le cerveau : les autres publié en juin 2017.

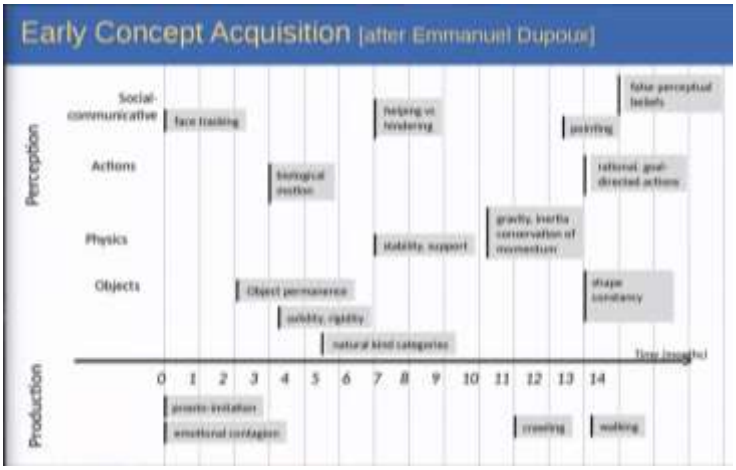
[302] Voir la critique très dure du HBP dans Dirty Rant About The Human Brain Project de Cathy O'Neil, 2015, qui note qu'il est bien trop tôt pour simuler le cerveau dans la mesure où l'on ne sait même pas décrire son fonctionnement correctement ainsi que Neuroscience: Where is the brain in the Human Brain Project? de Yves Frégnac et Gilles Laurent, 2014, qui s'interrogent sur la gouvernance du projet. J'ajouterai que la dernière conférence du HBP, le Human Brain Project Summit d'octobre 2018 n'est pas un bon gage de transparence : même son agenda est confidentiel !

[303] Source : Government Seeks High-Fidelity "Brain-Computer" Interface, février 2016.

De son côté, le projet **Allen Brain Atlas** planche sur la cartographie du cerveau de différentes espèces dont l'homme et la souris, au niveau de l'expression des gènes de ses différentes cellules nerveuses. La plateforme et les données associées sont ouvertes. Des chercheurs de l'Université de Berkeley ont même réussi à créer une cartographie précise de la sémantique du cortex[304].

Reste aussi, côté neurobiologie, à comprendre le processus d'apprentissage extraordinaire des enfants en bas âge et jusqu'à 20 ans. Comment le cerveau se câble-t-il pendant les phases d'apprentissage ? Comment séparer l'inné de l'acquis dans les processus d'apprentissage ?

Le slide *ci-contre* illustre les phases d'apprentissage d'un bébé et est issu de la conférence de Yann Le Cun à l'USI à Paris en juin 2018 (vidéo)[305].



On dissèque les souris, mais bien évidemment pas les enfants en bas âge. Donc, on ne sait pas trop. Et l'IRM est insuffisante. Les chinois et les japonais planchent sur une voie intermédiaire en cartographiant le cerveau de singes qui sont plus proches de l'homme que les rongeurs.

Pour résumer, un bon nombre de recherches portent sur le fonctionnement du cerveau, avec une intersection avec les recherches en intelligence artificielle.

Hacker le cerveau

Une fois que l'on a compris le fonctionnement du cerveau, on peut envisager de le hacker. Soit pour le réparer, soit pour l'améliorer. Les réparations sont relatives au traitement de certaines pathologies neurodégénératives comme la maladie de Parkinson, qui est déjà en partie traitable par un hack utilisant une électrode implantée dans le cerveau. L'amélioration passerait par la création d'implants cérébraux permettant de connecter de manière bidirectionnelle le cerveau à des IA et au cloud.

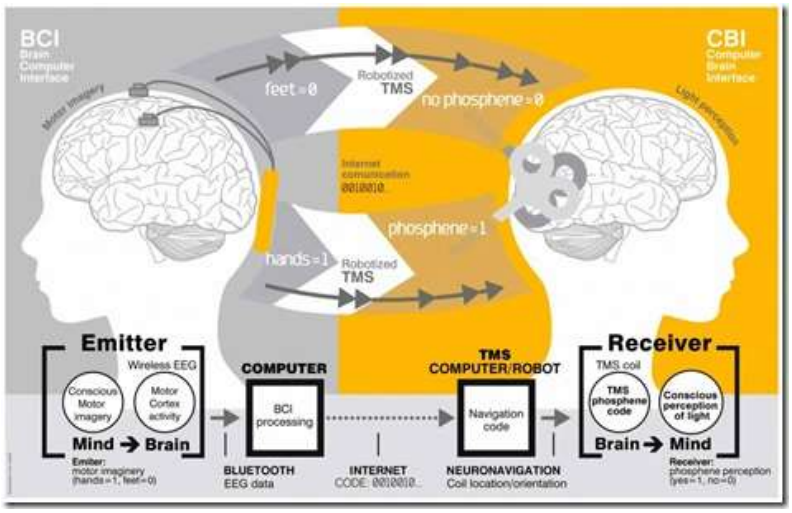
Dans "The age of spiritual machines" de 1999, Ray Kurzweil prévoyait qu'en 2029, des nanorobots connecteront notre cortex cérébral au cloud. Or Kurzweil s'est régulièrement planté sur ses prévisions concernant les nano-technologies.

Il oubliait quelques petits détails pratiques de bon sens : comment créer des nano-robots pour chaque neurone du néocortex qui capteraient et influenceraient leur état, et le transmettraient à l'extérieur ? Il y a 16 milliards de neurones à cet endroit et 2000 fois plus de synapses au minimum. Ca fait beaucoup d'informations à collecter et à émettre à l'extérieur du cerveau ! Si l'on s'appuyait sur des émetteurs radiofréquences, il serait intéressant d'évaluer l'énergie à consommer au sein du cerveau pour rendre cette communication possible, quand bien même celle-ci serait fournie par le glucose circulant dans le cerveau. On peut imaginer des nano-robots moléculaires. Mais aucun laboratoire n'a créé d'émetteurs d'ondes radio moléculaires. Ray Kurzweil se trompe au minimum sur le délai de la mise en œuvre de sa prévision. Au mieux aura-t-on à cet horizon des prothèses qui relieront les entrées/sorties du cerveau à l'extérieur (vue, cortex moteur). Pas le néocortex directement.

[304] Voir UC Berkeley team builds 'semantic atlas' of the human brain de Jessica Hall, 2016.

[305] Il précisait que les nombres de mois étaient mal alignés avec les fonctions apprises.

Certains se lancent dans la connexion avec le cortex cérébral cognitif et visuel, et pas seulement moteur. Des **expériences de télépathie** sont possibles, en captant par EEG la pensée d'un mot d'une personne et en la transmettant à distance à une autre personne en lui présentant ce mot sous forme de flash visuel par le procédé TMS, de stimulation magnétique transcrâniale[306]. Mais on transmet un seul « bit » avec ce genre de technique.



Un peu comme avec des casques de captation d'électro-encéphalogrammes qui permettent de piloter par concentration un seul paramètre mécanique d'un jeu. Si on peut déjà alimenter le cerveau au niveau de ses sens, comme de la vue, en interceptant le nerf optique et en simulant le fonctionnement de la rétine ou par la TMS, on ne sait pas l'alimenter en **idées** et **informations abstraites** car on ne sait pas encore vraiment comment et surtout où elles sont stockées. En tout cas pas encore car c'est l'ambition de startups américaines que d'y arriver un jour.

C'est le projet de **Neuralink**, une startup créée fin 2016 par Elon Musk et de **OpnWtr**[307] avec son bonnet utilisant des capteurs photos et des émetteurs infrarouges pour cartographier finement l'état des neurones et même, à terme, le modifier. **Facebook** essaye aussi de lire dans les pensées pour remplacer les claviers[308] ! Ces projets ont des niveaux d'ambition très différents. Elon Musk n'étant jamais à une exagération près, il prévoit que sa technologie de nano-électrodes permettra de lire et écrire dans les neurones à l'échelle individuelle pour remplir la mémoire.

C'est pour lui un moyen de relier l'Homme à l'IA pour qu'il puisse en tirer le meilleur parti et la contrôler. À ceci près que l'on n'a aucune idée de la manière dont est stockée la mémoire dans le cerveau et en particulier, que l'on ne sait pas relier sémantiquement chaque neurone du cortex à votre mémoire[309]. C'est au passage négliger un point clé : la capacité d'une IA maléfique de hacker le cerveau du porteur d'électrodes. Et si celles-ci ne peuvent être facilement enlevées ou débranchées, quel sera le recours de l'utilisateur en cas de problème ? Bref, cette solution est potentiellement pire que le problème qu'elle est censée résoudre.

---

[306] Voir Scientists pull speech directly from the brain par Devin Coldewey, Techcrunch, avril 2019. ainsi que pour lire dans les pensées, un petit code en Keras associé à un casque d'EEG : Using deep learning to "read your thoughts" — with Keras and EEG par Justin Alvey, mars 2019.

[307] J'ai publié en juin 2017 une étude détaillée des projets de Neuralink et OpenWatr dans une série de trois articles : http://www.oezratty.net/wordpress/2017/startups-bidouille-cerveau-neuralink/.

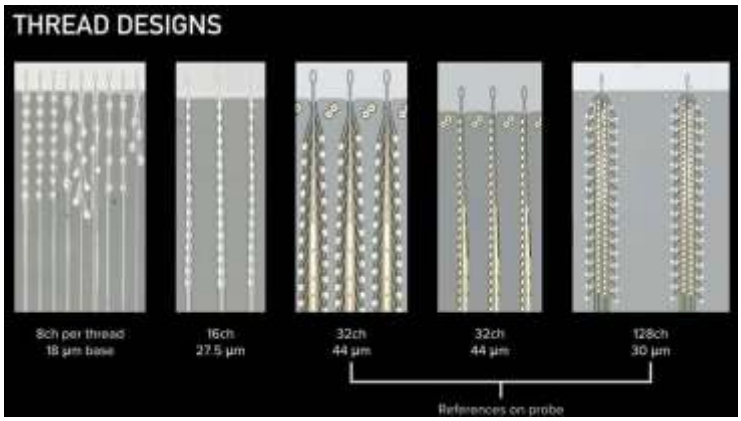
[308] Nanalyze a identifié 29 startups qui s'attaquent à l'interface cerveau-machine dans 29 Neurotech Companies Interfacing With Your Brain, octobre 2017.

[309] Cela n'empêche pas certains de prévoir que l'on pourrait remplir la mémoire directement avec ds électrodes. Voir Google "Brain Implants" Could Make Learning Obsolete in 20 Years, Says AI Expert par Loukia Papadopoulos, mai 2019.

Mais Elon Musk en juillet 2019 tentait de rassurer son monde : son système d'électrodes sera connecté via un récepteur sans fil qu'il suffira de désactiver en cas de problème[310]. Avec l'équipe de Neuralink, il levait le voile sur leur système. J'ai eu l'occasion de détailler le contenu de leur annonce[311] : en gros, ce sont des électrodes miniaturisées intégrées dans des fils souples de plusieurs formes.

Ces fils font un quart de l'épaisseur d'un cheveu et ont la taille d'un neurone. Ils ont créé un prototype de 96 fils comprenant 3072 électrodes, donc 32 électrodes par fil.

Les fils seraient posés dans le néocortex, en périphérie du cerveau, par un robot. Le tout permettrait par exemple de reconnaître les lettres auxquelles l'utilisateur pense pour les faire taper dans un clavier virtuel.



Les efforts de Neuralink n'aboutiront pas au point de permettre d'implanter des souvenirs dans le cerveau, mais qu'ils permettront tout de même probablement de faire des progrès dans le traitement de certaines maladies neurodégénératives qui nécessitent de stimuler certaines parties précises du cerveau, surtout limbique (hippocampe, thalamus, hypothalamus) avec ce que l'on appelle la stimulation cérébrale profonde[312]. Par contre, il vaut mieux éviter de gober les balivernes d'Elon Musk lorsqu'il déclare que sa solution permettra de traiter l'autisme[313] !

C'est d'ailleurs ce qu'aurait déjà réalisé un chercheur de l'**University of Southern California**, **Dong Song**, dans une expérience publiée fin 2017[314] avec sa solution pour améliorer la mémoire court terme de 15% à 30% (de quoi… ?[315]). Il s'agit d'un implant cérébral qui stimule électriquement l'hippocampe qui régit le fonctionnement de cette mémoire à court terme. La stimulation électrique imite celle qui intervient dans les personnes saines. C'est censé servir aux personnes atteintes de la maladie d'Alzheimer ou de démence. Mais la stimulation est indiscriminée. Elle n'agit pas sur des neurones individuels ou sur des souvenirs précis. D'autres chercheurs, de l'**Université de Boston**, Robert Reinhart et John Nguyen, ont fait de même en 2019 mais avec un système de stimulation non invasif exploitant un casque à électrodes[316].

C'est aussi ce qu'ambitionne de réaliser un certain Newton Howard d'Oxford avec l'implant neuronal Kiwi qu'il développe dans **Nitoo** (France). C'est une petite électrode qui capte des informations provenant des neurones ou agit dessus. La société est hébergée à Paris à l'Institut du Cerveau et de la Moelle Épinière et a été cofinancée par Bpifrance. La puce doit réduire les effets de maladies neurodégénératives comme celle d'Alzheimer ou de Parkinson.

---

[310] Voir Neuralink's Technology Is Impressive. Is It Ethical? Par Andrew Maynard, juillet 2019. Voir toutes les questions que posent ces systèmes de BCI dans Towards new human rights in the age of neuroscience and neurotechnology par Marcello Ienca et Roberto Andorno, 2017.

[311] Voir Ce que prépare Neuralink, Olivier Ezratty, juillet 2019.

[312] Voir Deep brain stimulation par Jonathan Pugh, 2019, qui décrit très bien l'histoire de la stimulation profonde du cerveau.

[313] Voir "This Could Be a Tragedy For Humanity" à propos de Neuralink par Gregg Braden, décembre 2019 (10 minutes) qui est alarmiste sur le potentiel de Neuralink, mais en exagérant ses capacités.

[314] Voir For the First Time Ever, Scientists Boosted Human Memory With a Brain Implant de Kristin Houser, novembre 2017.

[315] En fait, il semble que l'amélioration de performance concerne les personnes dont la mémoire court terme est défectueuse. Cela ne concerne pas les personnes en pleine forme.

[316] Voir Working memory revived in older adults by synchronizing rhythmic brain circuits, Nature, 2019.

Dans la lignée de Neuralink, nous avons aussi **Kernel** (2016, USA, $100M) qui souhaite créer une prothèse neuronale pour l'Homme[317] et plutôt dans la tendance « réparer l'Homme diminué ».

Il faut évidemment prendre avec des pincettes tous ces effets d'annonces. Ainsi, dans **Mashable**, une certaine Marine Benoit affirmait un peu rapidement en mars 2016 qu'une équipe avait mis au point "*un stimulateur capable d'alimenter directement le cerveau humain en informations*". À ceci près que l'étude en question, **Frontiers in Human Neuroscience** ne faisait état que d'un système qui modulait la capacité d'acquisition par stimulation ! Il en va de même pour d'autres travaux du même genre des HRL Labs, montés en épingle en 2020[318].

Pour l'instant, on doit se contenter de lire dans le cerveau dans la dimension mécanique mais pas "écrire" dedans directement. On ne peut passer que par les "entrées/sorties", à savoir les nerfs qui véhiculent les sens, mais pas écrire directement dans la mémoire.

Chez **OpnWtr**, l'idée est d'utiliser leur bonnet pour réaliser des IRM fonctionnelles captant l'activité du cerveau à une résolution faible, de l'ordre du millimètre cube. C'est un outil de démocratisation des IRM qui aujourd'hui s'appuient sur des infrastructures lourdes en environnement hospitalier. Le projet d'OpnWtr est d'inverser à terme le processus et d'utiliser la technologie qui s'appuie sur des rayons infrarouge pour agir sur les neurones. Mais au vu de la résolution, l'impact sera très macro.

Facebook, mène un projet pour capter les lettres et chiffres auxquels on pense avec un casque d'EEG (électro-encéphalogramme) et permettre une saisie plus rapide d'informations. C'est plausible d'un point de vue technique car une IRM fonctionnelle préalable permet d'identifier les zones associées dans le cortex. A ceci près qu'il n'est pas évident que cela permette une saisie de texte vraiment plus rapide qu'avec le doigt, un clavier ou la parole.

En septembre 2019, Facebook faisait l'acquisition, pour un montant non précisé mais élevé, de la startup **CTRL-labs** (2015, USA, $67M) qui développe un bracelet permettant de contrôler un système informatique à partir des micro-impulsions de l'utilisateur sur son bras. L'intérêt de leur système est qu'il est non invasif contrairement à ce que développe Neuralink.

Ce projet est voisin de celui du **MIT** qui a pu capter via un EEG les mots auxquels on pensait[319]. Un réseau de neurones permet de faire la correspondance entre l'EEG capté et les mots auxquels on pense. Il doit probablement être entraîné pour chaque utilisateur.

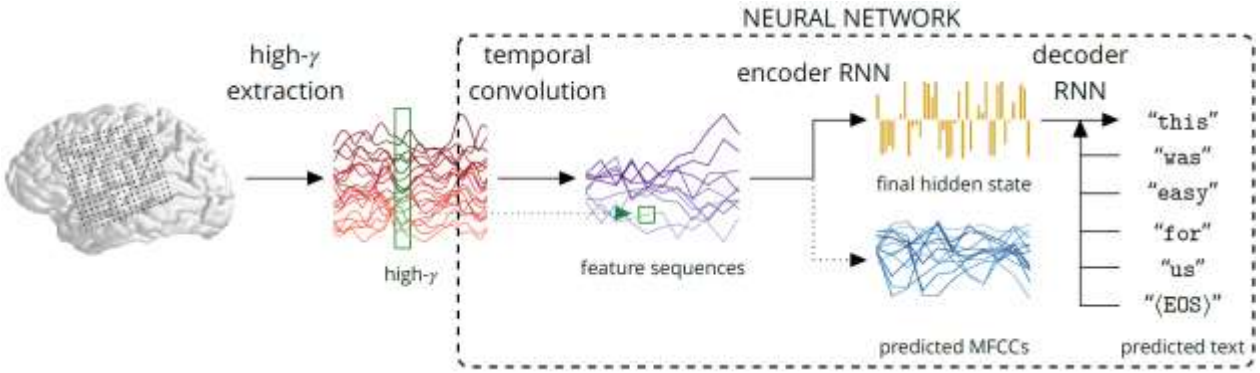
Ce genre d'outil passe par l'exploitation d'ECG multipoints (250 électrodes) placés dans le cortex de patients épileptiques complété par un réseau de neurones profond entraîné en leur faisant lire des phrases à haute voix. Le réseau est alors capable d'identifier les corrélations entre les mots et les zones du cerveau actives lors de la diction.

---

[317] Voir Kernel's Quest to Enhance Human Intelligence de Bryan Johnson, octobre 2016.

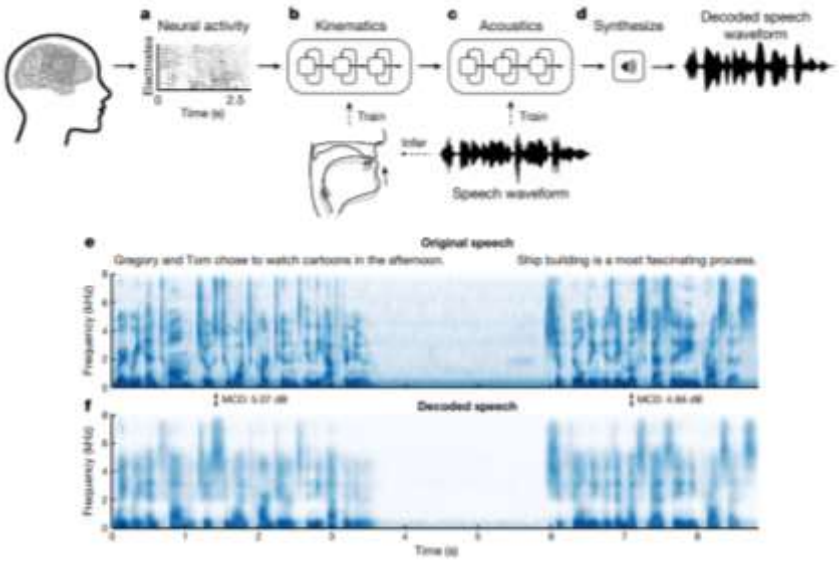
[318] Voir Scientists discover how to 'upload knowledge to your brain' qui relève d'une stimulation électrique du cerveau permettant d'améliorer l'apprentissage. Mais celui-ci a bien lieu physiquement et réellement. Le cerveau n'est pas alimenté artificiellement de l'extérieur. Voir Transcranial Direct Current Stimulation Modulates Neuronal Activity and Learning in Pilot Training par Jaehoon Choe et al, 2016.

[319] Voir Computer system transcribes words users "speak silently, avril 2018 et la vidéo associée.



Des chercheurs de l'**Université de Californie San Francisco** sont allés encore plus loin en 2019 en générant une voix de synthèse à partir des pensées de patients atteints de troubles moteurs comme l'épilepsie[320]. Le système reposait sur l'enregistrement de signaux électriques à l'aide d'un implant d'électrodes placées dans le cerveau (électrocorticographie ou ECoG) qui rappelle ce que fait Neuralink. Les signaux étaient enregistrés pour cinq patients en train de parler (vidéo).

Différents réseaux de neurones étaient alors utilisés pour exploiter les signaux électriques collectés et les relier aux paroles des patients. Un réseau de neurones à mémoire bLSTM analysait la représentation du langage dans les zones motrices du cerveau. Un autre réseau associait ces informations aux phonèmes générés par les locuteurs. Au bout du compte le réseau devenait capable de générer un son synthétique en lisant dans les pensées de l'utilisateur.



Un groupe de chercheurs chinois a sinon réussi à connecter un cerveau humain à celui de rats pour contrôler leurs mouvements[321]. Fascinant ! L'expérience utilise des électrodes implantées dans le cortex moteur du rat. La commande provenait de la détection de signaux d'électro-encéphalogramme humain.

Le programme **N3** (Next-Generation Nonsurgical Neurotechnology) de la DARPA lancé en 2019 vise à développer des interfaces homme machine non invasives comme des casques. L'idée est d'utiliser cela pour commander plus efficacement des drones, des claviers. Des projets ont reçu pour ce faire entre $18M et $20M de financement. L'objectif est de pouvoir lire et écrire dans des zones du cerveau d'un millimètre cube.

[320] Voir New brain-machine interface can generate natural-sounding synthetic speech using brain activity par Ibezim Chukwuemerie, avril 2019 qui fait référence à Speech synthesis from neural decoding of spoken sentences par Gopala Anumanchipalli & al, Nature, avril 2019 (20 pages). Voir L'activité cérébrale retranscrite en mots grâce à cette IA par Edward Back, avril 2020 ou Une intelligence artificielle lit la parole dans le cerveau par Elena Sender, juillet 2020 qui font tous deux référence à Machine translation of cortical activity to text with an encoder–decoder framework par Joseph G. Maki et al, juin 2019 (22 pages) d'où vient l'illustration.

[321] Voir Human Mind Control of Rat Cyborg's Continuous Locomotion with Wireless Brain-to-Brain Interface, février 2019 (12 pages).

C'est moins impressionnant pour la lecture que ce que fera l'IRM NeuroSpin en France avec sa résolution d'un dixième de millimètre. Les techniques utilisées pourront être à base de guides de lumière, d'ultrasons, d'interférences électromagnétiques, ou de particules magnétoélectriques implantées dans le cerveau[322].

Copier le cerveau

Après le hacking du cerveau, pourquoi ne pas copier directement son contenu dans une machine et nous faire vivre une vie virtuelle à la Matrix ? C'est plus *border line* d'un point de vue technologique mais cela n'empêche pas certains d'y penser et d'y travailler.

Dans "The Singularity is Near[323]", Ray Kurzweil fantasmait sur la capacité à venir de transplanter un cerveau dans une machine et d'atteindre ainsi l'immortalité, incarnation ultime du solutionnisme technologique qui cherche à trouver une solution technologique à tous les problèmes ou fantasmes humains.

Le *dump* du contenu d'un cerveau dans un ordinateur fait cependant face à quelques obstacles technologiques de taille. Heureusement d'ailleurs[324] !

Quels sont-ils ? Tout d'abord, on ne sait pas encore précisément décrire le mode de stockage de l'information dans le cerveau. Se situe-t-il dans les neurones ou dans les synapses qui relient les neurones aux axones d'autres neurones ? Dans Memories may not live in neurons synapses paru dans Scientific American en 2015, il est fait état que l'information serait stockée dans les neurones et pas au niveau des synapses[325].

Ce stockage est-il du même ordre dans le cortex et dans le cervelet ? Qu'en est-il du cerveau limbique qui gère les émotions, le bonheur et la peur, en interagissant à la fois avec le cortex et avec les organes producteurs d'hormones ? On cherche encore !

L'information est probablement stockée sous forme de gradients chimiques et ioniques, probablement pas sous forme binaire ("on" ou "off") mais avec des niveaux intermédiaires. En langage informatique, on dirait que les neurones stockent peut-être des nombres entiers voire flottants au lieu de bits individuels. Il n'est pas exclu non plus que les neurones puissent stocker plusieurs informations à différents endroits (dendrites, synapses, axones, microtubules à l'intérieur des cellules) sans compter l'hypothèse « quantique » qui verrait les neurones stocker de grosses quantités d'informations dans les microtubules.

La communication entre deux neurones est chimique, via un potentiel d'ions calcium, sodium et potassium, et régulée par des hormones de régulation de la transmission nerveuse telles que l'acétylcholine, la dopamine, l'adrénaline ou des acides aminés comme le glutamate ou le GABA (acide γ-aminobutyrique) qui bloquent ou favorisent la transmission d'influx nerveux.

A cette complexité, il faut ajouter l'état des cellules gliales qui régulent l'ensemble et conditionnent notamment la performance des axones via la myéline qui l'entoure. La quantité de myéline autour des axones est variable d'un endroit à l'autre du cerveau et module à la fois l'intensité et la rapidité des transmissions nerveuses. Cela fait une complexité de plus dans le fonctionnement du cerveau !

Et si la mémoire n'était constituée que de règles et méthodes de rapprochement ? Et si le savoir était en fait encodé à la fois dans les neurones et dans les liaisons entre les neurones ? En tout cas, le cerveau est un gigantesque puzzle chimique qui se reconfigure en permanence.

---

[322] Voir DARPA Funds Ambitious Brain Machine Interface Program par Megan Scudellari, mai 2019.

[323] L'ouvrage est librement téléchargeable dans The Singularity is near.

[324] Voir The empty brain de Robert Epstein, mai 2016, pour qui ce n'est même pas la peine d'essayer !

[325] Découverte confirmée par des chercheurs du MIT début 2016. Voir MIT discovers the location of memories: Individual neurons, mars 2012.

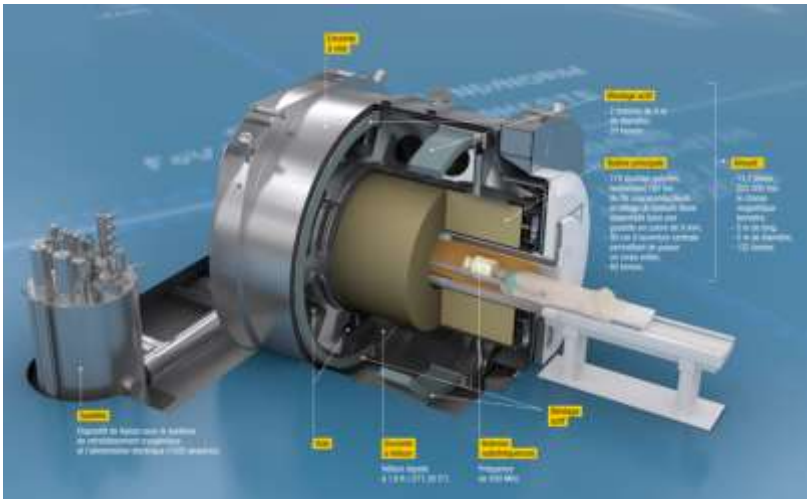
Les neurones ne se reproduisent pas mais leurs connexions et la soupe biologique dans laquelle ils baignent évoluent sans cesse.

Comment détecter ces potentiels chimiques qui se trouvent à des trillions d'endroits dans le cerveau, soit au sein des neurones, soit dans les liaisons interneuronales ? Comment le faire avec un système d'analyse non destructif et non invasif ?

Il n'y a pas 36 solutions : il faut passer par des ondes électromagnétiques, et avec une précision de l'échelle du nanomètre. Aujourd'hui les scanners utilisent généralement trois technologiques : la tomodensitographie qui mesure la densité de la matière par rayons X, les PET scanners qui détectent des traceurs biologiques radioactifs par émission de photons et l'IRM qui détecte les corps mous par résonnance magnétique nucléaire, qui n'irradie pas le cerveau mais doit le plonger dans un bain magnétique intense. Ces scanners ont une résolution qui ne dépasse pas l'ordre du millimètre et elle ne progresse pas du tout en suivant une loi exponentielle de Moore !

Le dernier système en cours de mise en place dans le laboratoire **NeuroSpin** du CEA à Saclay, que nous avons déjà cité.

Il s'agit du système franco-allemand **Iseult**, le scanner d'IRM corporel le plus puissant du monde, équipé d'un aimant record de 11,7 Teslas et 132 tonnes. Son bobinage supraconducteur en niobium-titane refroidi par cryogénisation à l'hélium pèse 45 tonnes (*ci-dessous*, source). Il complétera l'IRM dotée d'un aimant de 7 tonnes qui est opérationnelle chez Neurospin depuis 2008. Plus l'aimant est puissant, plus on augmente la résolution de l'IRM[326].



Ce système va servir à générer des images 3D de plus haute résolution, descendant en-dessous du mm3 de l'IRM traditionnelle. Elle descendrait au niveau du dixième de mm (100 microns). Il est pour l'instant difficile d'aller en-deçà avec des techniques non invasives.

Iseult permettra d'identifier plusieurs types de molécules au-delà de l'eau, comme le glucose ou divers neurotransmetteurs, notamment via l'injection de marqueurs à base de molécules magnétisées. La mise en service prévue initialement pour 2018 est en retard de plusieurs années sur le calendrier initial. A terme, on pourra aller jusqu'à observer le fonctionnement des neurones individuellement.

Ce projet rappelle qu'une autre exponentielle est en jeu : plus on veut observer l'infiniment petit, plus l'instrument est grand et cher. Comme pour les accélérateurs de particule et le LHC du CERN pour la découverte du boson de Higgs. Plus on augmente la résolution de l'IRM fonctionnelle, plus il faut augmenter la fréquence de scan et la puissance de l'aimant, donc sa taille.

[326] L'aimant a été conçu avec le concours du CEA-Irfu, l'Institut de Recherche sur les lois Fondamentales de l'Univers, qui a réutilisé ses acquis issus de la création des aimants supraconducteurs du Large Hadrons Collider du CERN de Genève. Il est fabriqué par Alstom-GE à Belfort, l'intégration du scanner étant réalisée par l'allemand Siemens, l'un des leaders mondiaux de l'IRM médicale. Y contribue également la société française Guerbet, spécialisée dans la production d'agents de contraste utilisés dans l'imagerie médicale.

D'où l'intérêt de la solution légère et, en apparence, très élégante, de OpnWatr évoquée plus haut mais qui n'a pas encore fait ses preuves, notamment en termes de résolution spatiale.

Des capteurs d'électro-encéphalogrammes existent bien (EEG). Ils sont placés à la périphérie du cortex sur la tête et captent l'activité de grandes zones de contrôle psychomotrices du cerveau avec un faible niveau de précision.

C'est très "macro". La mémoire et le raisonnement fonctionnent au niveau du "pico". Qui plus est, si on sait cartographier approximativement les zones fonctionnelles du cerveau, on est bien incapable de capter le rôle de chaque neurone et connexions neuronales prises individuellement.

Pourra-t-on connaître avec précision la position et l'état de toutes les synapses dans l'ensemble du cerveau et à quels neurones elles appartiennent ? Pas évident ! Autre solution : cartographier le cortex pour identifier les patterns de pensée. Si on pense à un objet d'un tel type, cela rend peut-être actif des macro-zones distinctes du cerveau que l'on pourrait reconnaître. Mais cela demanderait un entraînement personnalisé laborieux et de passer des heures allongé dans un scanner d'IRM.

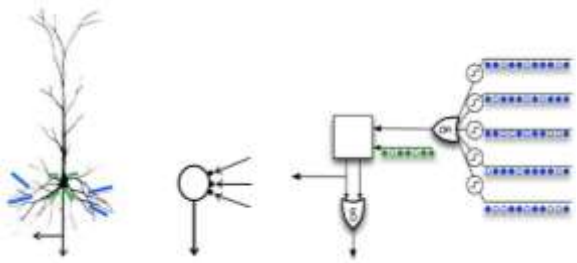
L'autre solution, imaginée dans le cadre du Human Brain Project consisterait à congeler notre cerveau puis à le découper en fines tranches qui seraient ensuite scannées au microscrope électronique. Cela supposerait que le patient soit en phase terminale d'un cancer affectant d'autres organes que le cerveau. C'est ce que l'on appelle une technique destructive[327] !

Recherche en AGI

L'AGI est un thème de recherche actif. Il intègre de la recherche à haut niveau conceptuel pour bâtir des systèmes intégrant les mécanismes de l'intelligence et à un niveau pratique avec des startups qui ambitionnent de révolutionner le secteur, souvent en associant les notions de deep learning et divers systèmes de gestion des connaissances. Peu de chercheurs affichent ouvertement vouloir créer une AGI. Ils sont en général très spécialisés dans un domaine précis de la gestion des connaissances et du raisonnement.

Dans la recherche fondamentale, on compte le projet **MICrONS** financé par l'IARPA, la DARPA du renseignement, lancé par George Church qui voulait carrément émuler le cerveau. Le projet a été mené entre 2014 et 2015 et ne semble pas avoir eu de suite[328].

Les tentatives affichées sont visibles chez quelques startups. Il y a d'abord **Numenta** (2005, USA), lancée par le créateur de Palm, Jeff Hawkins, ce dernier ayant en 2004 publié l'ouvrage **On Intelligence**, où il tente de décrire le fonctionnement du cerveau et la manière de l'émuler[329].



Jeff Hawkins pense que le cerveau est principalement une machine prédictive qui n'est pas forcément dotée d'une capacité de calcul parallèle intensive mais plutôt d'une mémoire associative rapidement accessible. Il insiste sur l'importance du temps dans les mécanismes de rétropropagation mise en œuvre dans les réseaux neuronaux uniquement dans les phases d'apprentissage. Alors que le cerveau bénéficie d'une mise à jour sensorielle permanente.

---

[327] C'est ce que veut proposer la startup Nectome (USA, $120K). Voir A startup is pitching a mind-uploading service that is "100 percent fatal", mars 2018. Le procédé consiste à se faire embaumer le cerveau à basse température avec un liquide qui en préserve la structure. Ce sera proposé un de ces jours à des patients atteints de maladies incurables. Juste au moment de leur décès. Tant que ce n'est pas un cancer du cerveau, cela peut donner l'espoir de revivre un jour à la Matrix dans un ordinateur. Mais les chances sont plutôt réduites que ces clients soient un jour véritablement satisfaits.

[328] Voir Machine Intelligence from Cortical Networks (MICrONS) ainsi que Hybrid bio-opto-electronics for AI avec une vidéo de Georges Church présentant le projet.

[329] L'ouvrage est téléchargeable gratuitement ici. Mais le lien n'est pas sécurisé.

Les thèses de Jeff Hawkins sont intéressantes et constituaient un pot-pourri des connaissances en neurosciences il y a 15 ans. Elles sont évidemment considérées comme un peu simplistes[330]. Hawkins oublie négligemment le rôle du cervelet et du cerveau limbique dans les apprentissages et le prédictif. Le cervelet contient plus de neurones que le cortex et il gère une bonne part des automatismes et mécanismes prédictifs, notamment moteurs.

Leur solution, Grok, permet de détecter des anomalies dans des systèmes industriels et informatiques. Elle imite le fonctionnement du cortex cérébral et de principes biologiques reprenant le principe de la mémoire par association et temporelle (Hierarchical Temporal Memory). Les réseaux de neurones à base de HTM utilisent des neurones plus sophistiqués que les réseaux de neurones habituels. Numenta propose aussi NuPIC (Numenta Platform for Intelligent Computing) sous la forme d'un projet open source[331].

En octobre 2018, Numenta annonçait une grande découverte sur le fonctionnement du cerveau sous l'appellation « The Thousand Brains Theory of Intelligence ». En gros, ils pensent avoir découvert comment fonctionnait la mémoire spatiale dans le néocortex au niveau des colonnes corticales[332]. Cela reposerait sur les *grid cells* qui capturent la position de notre corps par rapport aux objets qui nous entourent. Les grid cells ont été découvertes en 2005. A ce stade, ces travaux permettraient d'avancer dans la conception d'AGI en s'appuyant sur du biomimétisme, et notamment pour la conception de robots capables de bien se mouvoir dans l'espace de manière autonome.

La filiale de Google **DeepMind** fait des avancées régulières dans le chemin tortueux qui mène à l'AGI en s'appuyant sur la notion d'apprentissage neuro-symbolique, qui fusionne les réseaux de neurone du deep learning et la gestion des connaissances[333].

Ils annoncent plusieurs fois par an des avancées dans le domaine. En 2016, ils dévoilaient **PathNet**, un réseau d'agents intégrant des réseaux de neurones capable d'identifier la meilleure combinaison de réseaux de neurones pour atteindre un objectif donné.

Ils publiaient en 2017 des travaux sur la mémoire associative et en 2018 sur la modélisation du raisonnement abstrait[334].

[330] Voir ces critiques : Book Review: On Intelligence by Jeff Hawkins de Jeff Kramer, 2013, On Biological and Digital Intelligence de Ben Goertzel, et Is the model for general AI from On Intelligence by Jeff Hawkins reasonable and is it possible to use it practically?, 2014.

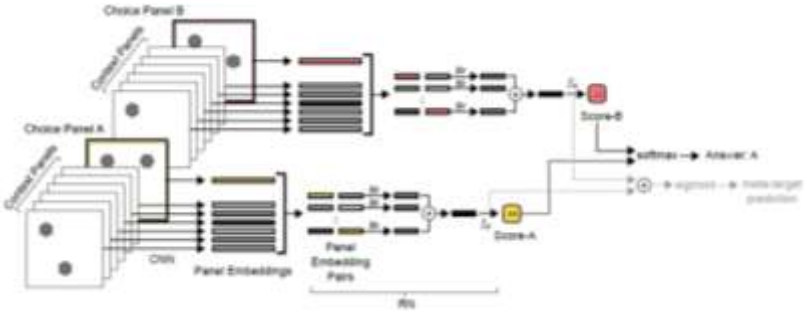
[331] L'intégration de la temporalité dans les réseaux de neurons est utilisée par d'autres chercheurs. Voir par exemple A New Type of AI Has Been Created Inspired by the Human Brain par Chris Young, août 2019, qui fait référence à Biological learning curves outperform existing ones in artificial intelligence algorithms par Herut Uzan & Al, de l'Université de Bar-Ilan en Israël, 2019 (11 pages).

[332] Voir A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex, octobre 2018 (15 pages) ainsi que Locations in the Neocortex A Theory of Sensorimotor Object Recognition Using Cortical Grid Cells, 2018 (16 pages). Et l'explication simplifiée dans Numenta Introduces Breakthrough Theory for Intelligence and Cortical Computation, octobre 2018.

[333] Voir Neurosymbolic Computation Thinking beyond Deep Learning, 2017 (43 slides), Neural-Symbolic Learning and Reasoning A Survey and Interpretation, 2017 (58 pages) et Toward Neural Symbolic Processing de Hang Li, 2017 (36 slides). Côté outils pour le neuro symbolique, il y a notamment la bibliothèque CILP++ lancée en 2015.

[334] En 2017, ils publiaient A simple neural network module for relational reasoning, 2017 (16 pages) dédié à la mémoire associative et en 2018, Measuring abstract reasoning in neural networks[334] (10 pages) pour modéliser du raisonnement abstrait de réponses à des tests élémentaires de QI.

Le tout, dans un dispositif comparant plusieurs types de réseaux de neurones (un réseau convolutionnel, un ResNet, un réseau à mémoire LSTM, et un réseau nouveau relationnel WReN, plus performant, *ci-contre*). Ce réseau de neurones est dédié à la résolution d'un seul problème particulier et relativement simpliste.

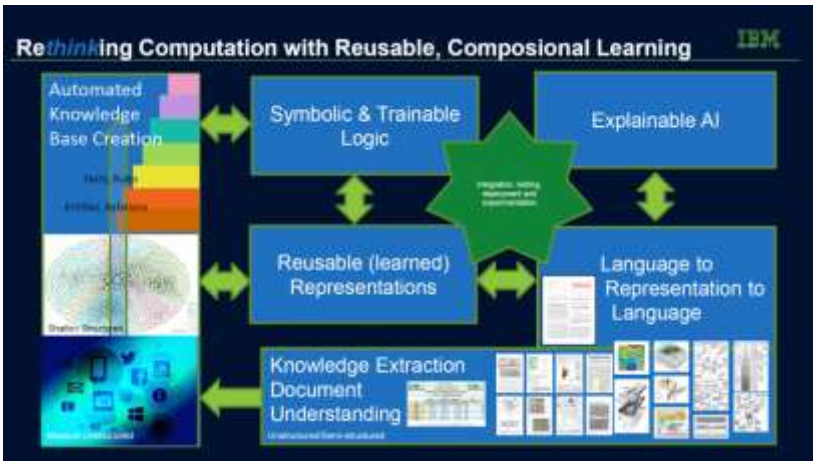


Peter Battaglia est un autre chercheur de DeepMind qui bâtit des modèles d'apprentissage prédictifs permettant de comprendre la structure des objets dans l'espace, leurs interdépendances et leurs mouvements[335]. Il utilise des réseaux de neurones de graphes qui sont des sortes de réseaux d'agents ou de SMA (Systèmes Multi-Agents). Le système expérimental permet de construire des structures stables mais on n'est pas encore au stade de l'AGI.

Nnaisense (2014, Suisse) a été créée par Jürgen Schmidhuber, le chercheur à l'origine des réseaux à mémoire LSTM utilisés dans le traitement du langage et la traduction. Ils ambitionnent de créer une AGI et dans un premier temps visent la finance, les industries lourdes et les véhicules autonomes. Comme ils travaillent en mode projet, ils ont déjà un chiffre d'affaires conséquent de $11M (2016). Mais ils sont encore très loins de l'ombre d'une AGI.

Des startups cherchent plus prosaïquement à faire coexister les approches symboliques à l'ancienne et connexionnistes. C'est par exemple le cas du Français **Nexyad** qui développe des solutions pour les véhicules autonomes, dont **SafetyNex** qui analyse les risques en temps réel en associant du deep learning – ne serait-ce que pour exploiter les contenus issus des capteurs – et du raisonnement ainsi que de la logique floue. Cela permet de créer une IA explicable, ce qui est critique pour la conduite de véhicules autonomes. Mais on n'est pas encore véritablement dans le champ de l'AGI car ces systèmes sont dédiés à des tâches très spécialisées.

Michael Witbrock, d'**IBM Watson** et ancien de CyCorp, décrit la démarche de fusion de l'IA symbolique et de l'IA connexionniste pour la gestion des connaissances. Le schéma *ci-contre* l'illustre avec en bas, l'extraction des connaissances par l'analyse du langage dans des documents[336]. Cela permet d'alimenter une base de connaissances (à gauche) elle-même exploitée par une IA symbolique (au milieu en haut).



Au passage, ce genre d'IA serait explicable.

La création d'AGI doit surmonter de nombreuses difficultés, la première étant le manque de formalisme et d'abstraction qui est utilisé pour décrire et décomposer les différentes composantes de l'intelligence humaine.

---

[335] Voir Structured Intelligence, de Peter Battaglia, Google DeepMind, 2019 (109 slides).

[336] Dans AI for Complex Situations: Beyond Uniform Problem Solving, 2017 (31 slides).

C'est la thèse de **Mike Loukides** qui souligne aussi le manque de rigueur scientifique de l'IA, qui relève le plus souvent de bricolage et d'une approche expérimentale pas toujours facilement reproductible manquant de fondements scientifiques, par exemple dans les mathématiques ou la philosophie[337]. Lorsqu'un domaine scientifique est bloqué, cela veut peut-être dire qu'il ne s'attaque pas aux bons problèmes. C'est peut-être le cas des approches qui cherchent à fusionner les moteurs de règles et le deep learning, ou les méthodes symboliques et connexionnistes.

L'autre écueil de l'IA et de l'AGI est que la mesure de son progrès est au mieux approximative. Comme la mesure est une abstraction d'une abstraction, si la première abstraction est erronée, la méta-abstraction le sera tout autant.

L'autre point lié à la créativité dans l'intelligence humaine est que celle-ci consiste souvent à casser des règles existantes voire de changer de mode d'abstraction[338]. Or les systèmes d'IA créatives à base de réseaux génératifs sont basés sur la combinaison de modèles et données existants. Ce qui génère un fonctionnement en boucle, comme lorsque des générateurs de musique essayent d'imiter le style des Beatles. Ces modèles n'auraient pas pu créer la musique des Beatles avec celle qui les précédait et les inspirait [339] !

David Chapman décrit la nature multidimensionnelle de l'IA qui associe des sciences, de l'ingénierie, des mathématiques, de la philosophie, du design et de la communication. Cela rend difficile la convergence autour de méthodes et approches communes[340]. L'IA est une discipline hybride et très expérimentale. Parfois, les méthodes formalisent trop vite les choses au lieu de rester dans le flou. Les travaux courants dans le deep learning génèrent des résultats Z avec des méthodes X et des données Y, parfois avec des progrès W par rapport à l'état de l'art, mais sans que l'on sache pourquoi ni que l'on puisse généraliser le résultat.

C'est par exemple le cas des Capsule Networks qui fonctionnent pour l'instant avec des lettres mais pas encore avec des images, sans que le passage de l'un à l'autre soit explicité. David Chapman pense que la recherche en IA pourrait être affectée des mêmes syndromes qui ont touché les sciences de la psychologie il y une décennie, dont les travaux étaient entachés de nombreuses fraudes et expériences difficiles à reproduire, ce qu'il appelle la crise de la réplication, aussi courante dans les biotechs[341].

Des conclusions erronées sont générées par les chercheurs lorsqu'ils communiquent sur leurs travaux en exagérant leurs effets, en interprétant les résultats de manière approximative, en ne relatant pas les essais infructueux de leurs recherches, en ne réalisant pas leurs expériences à une échelle suffisante, notamment d'un point de vue statistique[342].

---

[337] Dans l'excellent Why we need to think differently about AI août 2018.

[338] Voir à ce sujet Why we need to create AI that thinks in ways that we can't even imagine par Kevin Kelly, mai 2017 ainsi que A philosopher argues that an AI can't be an artist par Sean Dorrance Kelly en février 2019.

[339] Dans une interview de septembre 2019 dans l'émission de Stephen Colbert, **Paul MacCartney** expliquait que l'origine de sa créativité provenait de l'accumulation d'une grande quantité de morcaux ou bouts de musiques très variés, notamment sous l'influence de son père. Mais cette accumulation ne suffit pas. Elle est combinée à un gros travail de trial & error et de bouclage avec ses propres émotions ou avec celles d'auditeurs.

[340] Dans le très long How should we evaluate progress in AI?, 2018.

[341] Voir We need to improve the accuracy of AI accuracy discussions de Danny Crichton, mars 2018

[342] Exemple de biais incroyable avec cette étude New Use of A.I. Accurately Detected Cancer 86 Percent of the Time qui relate une IA qui détecte avec exactitude 86% de cancers colorectaux sur un échantillon de patients dont on savait déjà qu'ils étaient atteints de cette pathologie, créant un biais statistique énorme.

Zachary Lipton propose quelques recettes pour rendre les publications scientifiques de l'IA plus propres en évitant de mélanger explications et spéculation, en décrivant l'origine des améliorations de performance obtenues, en évitant de noyer le lecteur dans un jargon mathématique cryptique et en étant bien précis sur les termes employés[343]. Histoire d'éviter de tomber dans un optimisme béat[344]. Heureusement, d'autres sont plus sages sur les prédictions, comme la chercheuse **Melanie Mitchell** qui rappelle que les prévisions sur l'émergence d'AGI se sont régulièrement heurtées au mur de la difficulté de la réaliser[345].

Il se trouve que l'IA présente des limites, ne serait-ce que dans la capacité à résoudre des problèmes trop complexes (au-delà de NP-Difficile dans les classes de complexité) et ceux qui sont indécidables ! Selon le théorème d'incomplétude de Gödel, un système logique ne peut être à la fois complet et cohérent.

Il contient donc souvent des propositions indécidables. Une AGI devrait donc assister la prise de décisions faiblement formalisées et non simulables. Elle ne pourra pas décider à notre place face à l'incertitude surtout lorsqu'il s'agit d'innover [346]! Il en va de même pour les prédictions farfelues comme celles de la timeline *ci-contre*.

Il faut donc nous résigner à accepter le monde dans la complexité et la grande variété des paramètres qui le gouvernent. On ne peut pas tout prédire !

Mike Loukides décrit les errements de la période GOFAI de l'IA, celle des raisonnements à base de moteurs de règles[347].

Les critiques sur le deep learning d'aujourd'hui ne manquent pas non plus. Selon **Gary Marcus**, les ConvNet (réseaux de neurones convolutifs) requièrent trop d'informations et de calculs[348]. Ils appliquent une force brute trop brute et consommatrice de ressources machine. Il voit comme défis au deep learning le besoin de mieux gérer les informations hiérarchiques surtout dans le langage, le besoin de mieux différencier corrélations et causalités dans les analyses, comme l'Homme et de « désiloiser » le deep learning vis-à-vis de l'IA symbolique. Comme Yann Le Cun[349], il pense qu'il faut surtout développer les formes d'apprentissage non supervisé et les modèles hybrides symboliques/neuronaux qui observent le monde et en créent des modèles prédictifs à base de sens commun.

---

[343] Dans Troubling Trends in Machine Learning Scholarship, par Zachary Lipton, juillet 2018.

[344] Comme ce que l'on peut lire dans In the Future, There Will Be No Limit to What AI Can Accomplish in Science de Peter Rejce, mars 2018, qui survend effrontément les capacités existantes et à venir de l'IA.

[345] Dans Opinion | Artificial Intelligence Hits the Barrier of Meaning, par Mélanie Mitchell, novembre 2018.

[346] Je m'inspire ici en le paraphrasant de Jean Staune dans « Les clés du futur », page 594.

[347] Dans Why we need to think differently about AI.

[348] On en trouve dans Deep Learning: Diminishing Returns? de Bernard Murphy, juillet 2018, dans Deep Learning: A Critical Appraisal de Gary Marcus, 2017 (27 pages) et dans son post dans Medium In defense of skepticism about deep learning, janvier 2018.

[349] Voir aussi Hinton, Le Cun, Bengio : la « conspiration » du deep learning, de Benoît Georges dans Les Echos, juillet 2018.

K. Eric Drexler propose le concept de Comprehensive AI Services (CAIS) avec un raisonnement un peu récursif qui part du principe que la R&D en IA peut s'automatiser elle-même[350]. Il propose une distinction intéressante entre l'intelligence (de l'enfant qui apprend) et la compétence (de l'expert). D'ailleurs, lorsque l'on parle d'AGI, on précise rarement à quel étalon on se réfère ? Au commun des mortels ou au savant ?

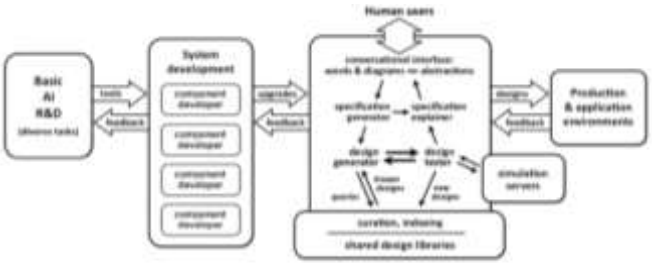


Figure 9: A task structure & architecture for interactive design engineering

Geoff Hinton pense qu'une approche connexionniste permettra de créer une AGI. Sur le fond, n'est pas absurde puisque le cerveau est aussi un réseau de neurones, certes avec des neurones plus sophistiqués que ceux que l'on simule dans des ordinateurs[351].

Pour **Yann Le Cun**, le problème du deep learning est que sa mémoire est limitée. D'où l'approche de Jeff Hawkins avec le HTM de Numenta et la mémoire temporale, même si cela n'a pour l'instant pas encore abouti. Il pense aussi que pour faire du raisonnement, il faudrait générer des résultats multiples et pas uniques avec des modèles basés sur les contraintes énergétiques, ce qui rappelle le concept des modèles génétiques. Bref, il faudrait combiner les approches neuronales et les approches par graphes. Et côté risque, Yann Le Cun botte en touche avec cette belle citation : *Will it want to take over the world? No, the desire to dominate is not correlated with intelligence but with testosterone*[352].

Rich Sutton estime de son côté que les progrès de l'IA viennent surtout de l'impact de la loi de Moore, indépendamment de la recherche de nouvelles méthodes et l'intégration formelle de connaissances structurées. Bref, comme s'il fallait se contenter de la force brute telle que celle qui est utilisée dans les jeux de société, des échecs au Go[353], combinée à de la simulation. L'analogie est aussi valable entre les réseaux de neurones convolutifs et les méthodes qui précédaient où l'on décrivait a la mano les formes que l'on recherchait dans les images.

La force brute a une fois de plus gagné. Le traitement du langage a l'air de générer le même phénomène, au moins au niveau des résultats, avec un deep learning probabiliste, même si la machine ne comprend pas ce qu'elle fait. Bref, il recommande aussi de désanthropomorphiser l'IA[354] !

Ce n'est pas encore le cas. La grande quête du moment dans les avancées de l'IA tourne autour de l'apprentissage non supervisé. Un apprentissage qui n'aurait pas besoin de gros volumes de données et qui apprendrait au fil de l'eau comme les Humains. En pratique, il est semi-supervisé car il sert surtout à détecter des anomalies ou nouvelles classes d'objets et comportements par rapport à ce qui est déjà connu et labellisé.

---

[350] Voir Reframing Superintelligence Comprehensive AI Services as General Intelligence par Eric Drexler, 2019 (210 pages).

[351] Voir Geoffrey Hinton, l'un des pionniers de l'IA, pense que le deep learning suffira à reproduire toute l'intelligence humaine : « je crois que l'apprentissage en profondeur va pouvoir tout faire », novembre 2020.

[352] Dans How Could Machines Learn Like Animals & Humans?, Yann Le Cun, conférence à Harvard, mars 2019 (134 slides).

[353] Dans The Bitter Lesson par Rich Sutton, mars 2019.

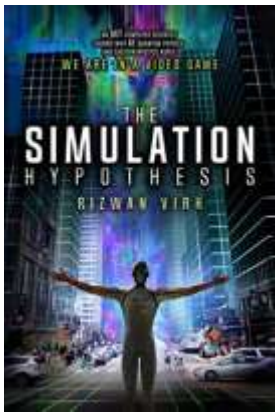
[354] Voici un exemple d'anthropomorphisation excessive d'une prouesse d'une IA dans Une intelligence artificielle suffisamment intelligente pour estimer ses propres erreurs par France Culture, novembre 2020 qui fait référence à Deep Evidential Regression par Alexander Amini et al, octobre 2019 (20 pages). Il s'agit ici d'un système de deep learning d'analyse d'image qui est capable d'évaluer l'incertitude statistique de ses inférences. Il ne raisonne pas.

Il y a une forte hybridation de l'apprentissage comme chez les humains qui associent leurs capacités d'observation, d'expérimentation et de raisonnement. On peut aussi appeler cela de l'apprentissage auto-supervisé (SSL pour self-supervised learning). L'expérimentation est de son côté associée à la notion d'apprentissage par renforcement. En pratique, les exemples d'apprentissage non supervisé sont toujours ad-hoc pour résoudre une tâche bien spécifique. Il n'existe pas pour l'instant d'apprentissage non supervisé générique. D'ailleurs, le non-supervisé est sémantiquement inexact car ces méthodes sont toujours supervisées à un titre ou à un autre[355].

Autre forme anthropomorphique, celle qui consiste à écrire que les cerveaux artificiels devraient aussi dormir, pour mettre en place des fonctionnalités qui rappellent les *garbage collectors* de certains langages de programmation mais qui s'appliquent en fait à des réseaux de neurones à impulsions ou pour simplemement introduire du bruit dans les réseaux de neurones[356].

Il faudrait peut-être même s'éloigner du cerveau pour créer des alter-intelligences plutôt que des intelligences tentant vainement d'imiter l'intelligence humaine[357]. La métaphore comparant l'oiseau et l'avion souvent utilisée par Yann Le Cun illustre ce propos et les limites de la bio inspiration. L'avion imite l'oiseau mais jusqu'à un point seulement. À la fin, l'oiseau est souple et peut se poser sur une cheminée alors que l'avion a besoin d'un aéroport. Mais il transporte des centaines de passagers, ce que ne fait pas l'oiseau. Il en va de même pour la force brute des outils numériques.

Il nous faut intégrer aux élucubrations du domaine la thèse selon laquelle nous vivons dans une simulation. Ceux qui l'avancent raisonnent plutôt au niveau philosophique. Lorsque l'on raisonne au niveau physique et théorie de l'information, cela devient rapidement moins plausible pour des raisons de complexité ultra-méga-exponentielle[358].



L'idée a été récemment relancée par un ouvrage de **Rizwan Virk**, un développeur de jeux vidéo du MIT, **The Simulation Hypothesis**. Mais l'approche n'est pas très scientifique. C'est de la science-fiction dans son plus pur état qui n'est pas très éloignée des explications du monde issues des grandes religions.

---

[355] Voir The Next Generation Of Artificial Intelligence et The Next Generation Of Artificial Intelligence (Part 2) par Rob Toews, 2020 qui fait un tour d'horizon des nouveautés du deep learning en 2020/2021 avec l'apprentissage non supervisé, le federated learning, les transformeurs, la compression des réseaux de neurones et les réseaux de neurones génératifs.

[356] Voir Artificial brains may need sleep too par Los Alamos National Laboratory, juin 2020 qui fait référence à Using Sinusoidally-Modulated Noise as a Surrogate for Slow-Wave Sleep to Accomplish Stable Unsupervised Dictionary Learning in a Spike-Based Sparse Coding Model par Yijing Watkins et al, 2020 (6 pages). Voir aussi Lack of Sleep Could Be a Problem for AIs par Garrett Kenyon dans Scientific American, décembre 2020 qui est contredit dans #5 AI Hype: AI could go psychotic due to lack of sleep!, décembre 2020.

[357] Voir Is the Brain a Useful Model for Artificial Intelligence? par Kelly Clancy, mai 2020.

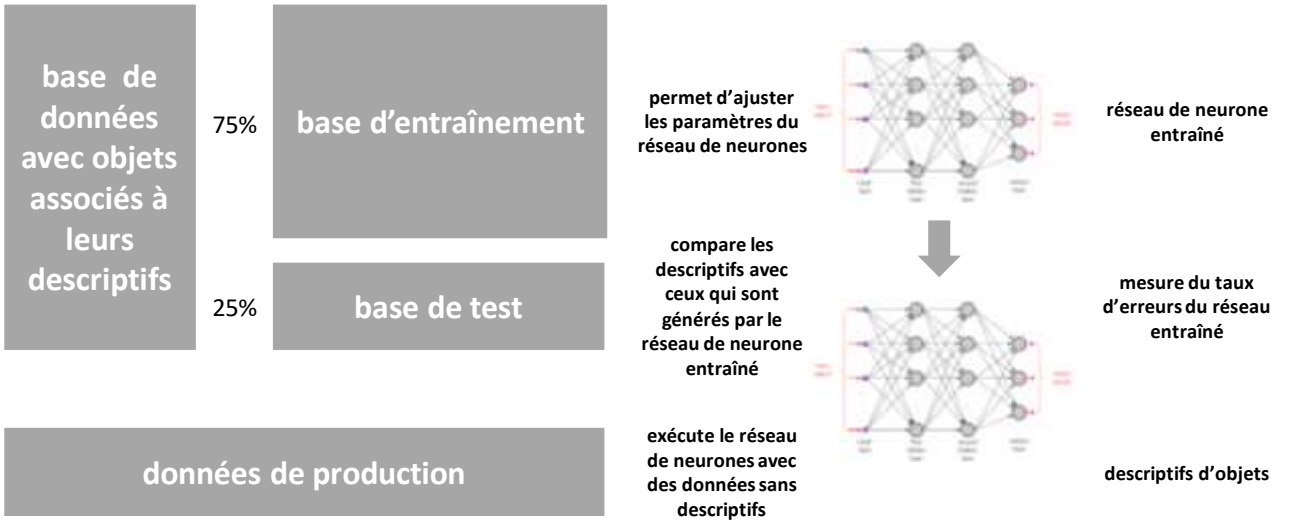
[358] Voir Are you living in a computer simulation? de Nick Bostrom, 2003 (15 pages), I, Sim - an Exploration of the Simulation Argument par Anders Hammarström, 2008 (47 pages), Pour un chercheur du MIT, il y a de fortes chances pour nous vivions dans une simulation, avril 2019 et Do We Live in a Simulation? Chances Are about 50–50 d'Anil Ananthaswamy, Scientific American, octobre 2020.

Données de l'IA

Le point commun du machine learning et de son sous-ensemble le deep learning est d'exploiter des données pour l'entraînement de modèles probabilistes. Avec les algorithmes/logiciels et le matériel, les données sont la troisième composante clé de la plupart des IA du jour.

La qualité des solutions d'IA est par conséquent étroitement liée à celle des données qui les alimentent. Avec l'IA et la big data, que nous traiterons plus tard dans ce document, la donnée est devenue une quasi-religion devant laquelle tout le monde se prosterne. On oublie cependant souvent que son accumulation ne tombe pas du ciel et est la conséquence de services de qualité ou d'infrastructures d'objets connectés et de télécommunications qui la collectent.

On distingue généralement trois types de données qui alimentent un système de machine learning : les données d'entraînement qui servent à paramétrer leur modèle, les données de test qui servent à valider le modèle et les données de production qui sont utilisées pour les faire tourner.



Pour des tâches de classification dans des modèles supervisés, les données d'entraînement et de tests contiennent leur label, à savoir, l'information qui doit être générée par le système à entraîner. C'est un jeu de test qui doit être doté d'une bonne répartition de l'espace du possible de l'application. On le découpe arbitrairement en deux sous-ensembles, l'un pour l'entraînement et l'autre pour les tests de qualification du réseau de neurones entraîné qui déterminent un taux d'erreur de reconnaissance. En général, la part de la base labellisée dédiée à l'entraînement est plus grande que celle qui est dédiée aux tests et dans un rapport 75% et 25%.

Les données d'entraînement et de tests sont indispensables pour la grande majorité des systèmes d'IA à base de machine learning, que ce soit pour de l'apprentissage supervisé ou non supervisé. L'apprentissage par renforcement utilise une plus faible quantité de données mais s'appuie en général sur des modèles déjà entraînés au préalable.

Données d'entraînement

Ce sont les jeux de données qui vont servir à entraîner un modèle de machine learning pour en ajuster les paramètres. Dans le cas de la reconnaissance d'images, il s'agira d'une base d'images avec le ou les labels qui en décrivent le contenu. Plus la base est grande, meilleur sera l'entraînement du système, mais plus il sera long. Si vous ne disposez pas de données déjà labellisées pour entraîner un modèle de machine learning ou de deep learning, vous n'irez pas bien loin ! Et la volumétrie nécessaire ? Elle est déterminée plutôt empiriquement et en fonction de l'état de l'art.

Les bases d'entraînement d'images ont une taille qui dépend de la diversité des objets à détecter. Dans l'imagerie médicale, des bases d'entraînement de pathologies spécialisées peuvent se contenter de quelques centaines de milliers d'images pour détecter quelques dizaines ou centaines de pathologies.

internes

- bases métiers
- trafic web & mobile
- objets connectés

externes

- open data publiques
- ImageNet, MNIST, WordNet
- internet
- réseaux sociaux

qualité et volumétrie

ImageNet : 14 millions d'images en open data

A l'autre extrémité de la complexité, la base d'entraînement d'images de Google Search s'appuie sur plusieurs centaines de millions d'images et permet la détection de plus de 20 000 objets différents.

L'entraînement d'un système de 50 000 images dure environ un quart d'heures avec des ressources dans le cloud mais le temps de calcul dépend bien entendu des ressources allouées comme le nombre de serveurs « compute », notamment de GPU Nvidia qui sont les plus courants. Lorsque l'on passe à des centaines de millions d'images, il faut des milliers de serveurs et jusqu'à plusieurs semaines pour l'entraînement !

Dans la pratique, les jeux d'entraînement de solutions de deep learning sont limités en taille à cause de la puissance de calcul nécessaire. On peut aussi réaliser un entraînement incrémental au gré de l'ajout de données avec des techniques de transfert de réseau de neurones[359].

Il est évidemment nécessaire de disposer de données d'entraînement de qualité, ce qui nécessite souvent un gros travail de filtrage, de nettoyage et dédoublonnage préalable à l'ingestion des données, une tâche existant déjà dans le cadre d'applications de big data.

Données de test

Ce sont les données, également labellisées, qui serviront à vérifier la qualité de l'entraînement d'un système. Ces données doivent avoir une distribution statistique voisine des données d'entraînement, au sens où elles doivent être bien représentatives de la diversité des données que l'on trouve dans la base d'entraînement et que l'on aura dans les données de production.

Les données de tests sont un sous-ensemble d'un jeu de départ dont une partie sert à l'entraînement et une autre partie, plus limitée, sert aux tests. Elles seront injectées dans le système entraîné et on en comparera les labels résultants avec les labels de la base. Cela permettra d'identifier le taux d'erreur du système. On passera à l'étape suivante lorsque le taux d'erreur sera considéré comme acceptable pour la mise en production de la solution.

Le niveau de taux d'erreur acceptable dépend de l'application. Son maximum généralement accepté est le taux d'erreur de la reconnaissance humaine. Mais comme on est généralement plus exigeant avec les machines, le taux véritablement accepté est très inférieur à celui de l'Homme, surtout par exemple pour des applications critiques comme dans le diagnostic médical.

La mesure du taux d'erreur ne doit pas être une simple moyenne. Elle doit intégrer un écart type et être capable de mettre en évidence les cas d'erreurs extrêmes qui pourraient avoir un impact négatif grave sur la solution et ses utilisateurs. Cela peut être par exemple les cas de faux négatifs dans les systèmes d'interprétation d'imagerie médicale ou encore les faux positifs dans le cas d'applications destinées à la police ou à la justice.

[359] Voir par exemple Incremental Learning in Deep Convolutional Neural Networks Using Partial Network Sharing, 2017 (12 pages).

Données de production

Il s'agit des données non labellisées qui alimentent le système lors de son utilisation en production. Le modèle de machine learning fournit alors le label manquant aux objets soumis.

Alors que les données d'entraînement sont normalement anonymisées pour l'entraînement du système, les données de production peuvent être nominatives ainsi que les prévisions associées générées par la solution.

La nouvelle réglementation RGPD de l'Union Européenne exige que les entreprises conservent les données personnelles des utilisateurs ainsi que les données générées. Cela concerne donc à priori les données générées par les systèmes à base d'IA. Une donnée personnelle générée artificiellement reste une donnée personnelle ! Et son origine artificielle doit être connue et traçable en cas d'audit.

Données de renforcement

J'utilise cette expression pour décrire les données qui servent à l'apprentissage par renforcement. Dans un chatbot, cela sera par exemple les données de réactivité des utilisateurs aux réponses des chatbots permettant d'identifier celles qui sont les plus appropriées.

En quelque sorte, ce sont des résultats d'A/B testing réalisés sur les comportements d'agents à base d'IA. Tout ce qui pourra être capté sur la réaction du monde réel et physique aux agissements d'un agent à base d'IA permettra potentiellement d'en ajuster le comportement par réentraînement.

L'apprentissage par renforcement est une sorte d'apprentissage supervisé incrémental car on l'utilise pour faire évoluer par petites touches impressionnistes des systèmes déjà entraînés.

Origine des données

Les données alimentant les systèmes d'IA proviennent de l'intérieur et/ou de l'extérieur de l'entreprise. Avec ça, on a bien couvert l'étendue du possible !

Elles sont issues de toutes sortes de capteurs divers : des objets connectés, du plus simple (thermomètre connecté) aux plus sophistiqués (machine outil, smartphone, ordinateur personnel). Comme pour les applications de big data habituelles, les sources de données doivent être fiables et les données bien extraites et préparées avant d'être injectées dans les systèmes à base de machine comme de deep learning.

Les solutions les plus avancées exploitent conjointement des données ouvertes externes et les croisent aux données que l'entreprise est seule à maîtriser. C'est un bon moyen de créer des solutions différenciées.

Les entreprises peuvent facilement croire que seuls les grands acteurs du marché tels les GAFA disposent de données en quantité suffisante pour entraîner leurs IA. Les entreprises disposent en fait de tombereaux de données exploitables, surtout dans la mesure où leur métier n'est pas assuré par les GAFA (banque, énergie, services publics, assurances, retail).

Qui plus est, il n'y a pas que les GAFA mais d'autres leaders du numérique grand public qui collectent des données. Ces acteurs le font cependant plutôt dans le grand public. Les données B2B sont pour l'essentiel dans les entreprises de marchés verticaux[360].

Les données ouvertes sont issues de l'open data gouvernementale[361], des réseaux sociaux et de différents sites spécialisés dans la fourniture de données, soit ouvertes, soit payantes, comme des bases de prospects d'entreprises ou de particuliers, selon les pays et législations en vigueur.

---

[360] J'explique tout cela dans Les GAFA, les entreprises et les données de l'IA, Olivier Ezratty, juillet 2019.

[361] Voir l'inventaire des données ouvertes publiques en France : data.gouv.fr Edition 2019, Tome 1 Le Best Of (904 pages).

Côté images, il y a par exemple la fameuse base de référence **ImageNet**, la base **LSUN** qui contient des points de vue d'extérieur, **CIFAR-10** et **CIFAR-100**, des bases d'images diverses basse résolution et **Celeba**, qui contient 200 000 photos de visages de célébrités. Dans le langage, il y a notamment la base lexicale **WordNet** (anglais) avec ses 117 000 expressions et **MNIST** (écriture manuscrite). On trouve des données publiques dans des secteurs plus spécifiques comme par exemple **Plant Village** (terrains agricoles), **Leafsnap** (une base de végétaux) ou encore **ImageCLEF** (la base des images de Wikipedia).

En France, le **Health Data Hub** publie depuis le printemps 2019 les données de santé ouvertes. Emmanuel Bacry, chercheur au CNRS en est le directeur scientifique. Il reprend les missions de l'**INDS** (Institut National des Données de Santé). L'agence **Etalab** recense sinon les jeux de données publics en open-data[362]. Les données de l'**IGN** sont maintenant ouvertes. La startup **nam.R** (2017, France) scrappe les données ouvertes françaises et mondiales et les met en forme pour les entreprises.

Le nettoyage puis la labellisation sont des défis majeurs de l'adoption rapide du machine learning par les entreprises[363]. Des sociétés spécialisées emploient des personnels plus ou moins spécialisés et plus ou moins bien rémunérés pour labelliser les données.

La préparation des données comporte de nombreuses tâches : supprimer les données erronées et les doublons, standardiser le format des données, compléter les informations manquantes, enrichir les données avec des sources d'informations complémentaires, anonymiser les données, supprimer les biais identifiés pouvant provenir d'un mauvais échantillonnage (par exemple, par catégories d'individus) et extraire des échantillons pour l'entraînement.

Les données d'entraînement des systèmes de machine learning à apprentissage supervisé doivent être bien labellisées, soit manuellement soit automatiquement, soit semi-automatiquement.

De nombreuses bases de référence d'images taggées l'ont été via de la main d'œuvre recrutée en ligne via des services du type d'**Amazon Mechanical Turk**[364] ou de **FigureEight** (2007, USA, $58M, aussi appelé CrowdFlower).

---

[362] Voir https://static.data.gouv.fr/resources/edition-papier-data-gouv-fr/20190401-091828/datagouv-edition-2019-tome1.pdf

[363] Voir This CEO is paying 600,000 strangers to help him build human-powered AI that's 'whole orders of magnitude better than Google', de Matt Weiberger, octobre 2018, et Could data costs kill your AI startup?, de Ivy Nguyen, novembre 2018.

[364] Voir Inside Amazon's clickworker platform: How half a million people are being paid pennies to train AI, de Hope Reese et Nick Heath, 2016.

Ces petites mains sont en quelque sorte les ouvriers de l'IA ! On peut aussi les trouver dans de nombreux sites d'intermédiation de travailleurs freelance comme **Crowd Guru** (2008, Allemagne), **CrowdSource** (USA), **fiverr** (2010, Israël, $111M), **foulefactory** (2014, France) et **clickworker** (2005, Allemagne, 10,7M€). Pour sa part, **Isahit** (2016, France, 2M€) est focalisée sur les tâches de labellisation pour le machine learning et ambitionne d'utiliser ce dernier pour faire la meilleure correspondance entre les missions et les freelances qui les exécutent, principalement situés en Afrique.

La startup **Vainu** (2013, Finlande) a eu de son côté l'idée de faire appel à des prisonniers pour nettoyer et labelliser ses bases de données d'entreprises à prospecter[365]. **Hive** (2013, USA, $20,2M) fait appel à plus de 700 000 utilisateurs d'applications mobiles pour labelliser des images. Il y a aussi ces fameux captcha qui serviraient à labelliser des images mais j'ai un doute au vu de leur grande spécialisation autour des feux rouges et autres éléments de signalétique routière. Dans le même ordre d'idée, **Scale AI** (2016, USA, $122M) labellise des images avec un mix de machine learning et de travail manuel distribué[366]. Ils ont Uber et OpenAI comme clients.

D'autres comme **thresher.io** (2015, USA), **OpenRefine** (open source) et **trifacta** (2012, USA, $124,3M) proposent des produits de collecte semi-automatique de labels afin de soulager les entreprises de la charge de labellisation. Cette labellisation peut cependant être de mauvaise qualité car, pour des applications industrielles, une expertise métier est souvent nécessaire à une labellisation correcte.

Picsell.ia (2020, France) est une startup Toulousaine qui propose un workflow collaboratif de labellisation d'image. Il détoure les objets dans les images et propose aux utilisateurs de labelliser les objets.

Le projet open source **Pixano** du CEA-List lancé en février 2020 permet l'annotation d'images par segmentation d'images au pixel près. Il assiste les opérateurs humains qui labellisent les objets détectés par segmentation.

Unbiased (2018, Suède) propose depuis 2020 son outil Data Marketplace, une solution intégrée de machine learning censée régler les problèmes éthiques grâce à une blockchain Telos. Le principe a l'air de relever du suivi des données qui circulent dans un processus allant de l'acquisition jusqu'à l'entraînement de modèles de machine learning et passant notamment par l'annotation ou labellisation des données. Leur Data Marketplace s'appuie sur l'Unbiased WorkForce, une plateforme de crowdsourcing de collecte de données et d'annotation d'images, textes et audio. C'est un service proposé sous forme d'abonnement. Une Blockchain gère le suivi de la réputation des contributeurs. La plateforme propose aussi de la gamification et du paiement pour les labelliseurs de données, un peu comme Figure8.

Cela étant, on a du mal à comprendre comment ce modèle évite les biais de données. Ceux-ci proviennent d'un grand nombre de paramètres qualitatifs qu'un système de traçabilité de ce genre n'a pas l'air de pouvoir éviter.

[365] Voir En Finlande, des prisonniers dressent des intelligences artificielles par Lila Meghraoua, mars 2019.

[366] Voir Silicon Valley's Latest Unicorn Is Run by a 22-Year-Old par Ashlee Vance, 2019.

Pour pallier ces problèmes, **Neuromation** (2017, USA, $25M), **Deep Vision Data** (USA), **Kubity** (2012, France) et **Oktal** (1989, France) proposent de générer automatiquement des données d'entraînement numériques d'objets 3D. **Zumo Labs** (2019, USA, $150K) génère des données d'entraînement synthétiques représentant les poses et le comportement humains.

L'offre Chameleon de **Mindtech** (2016, UK) permet de générer des vidéos labellisées avec des images bien segmentées à partir d'un éditeur de scénarios 3D servant à entraîner des systèmes de vision artificielle, notamment pour la conduite automone et la vidéosurveillance dans les smart cities. *Ci-contre*, voici un exemple de simulation 3D de situation consécutive à d'accident sur la voirie en ville et de nuit.

Une autre méthode consiste à utiliser des réseaux de neurones génératifs de type GAN pour générer des objets prélabellisés servant ensuite à entraîner un classifieur d'images. C'est notamment utile pour créer des vues d'objets 3D en 2D sous plusieurs angles de vue pour entraîner des réseaux convolutifs de vision[367]. Les cas d'usage sont cependant restreints[368]. Certains experts estiment que cette friction constitue une motivation supplémentaire à l'extension des usages du deep learning et de l'apprentissage par renforcement[369].

En dernier ressort, si une version numérique 3D de l'objet n'est pas disponible, on peut utiliser des automates qui vont le photographier sous tous les angles et avec des variations d'éclairage pour recréer des situations variées. Les photos captées alimenteront alors la base d'entraînement du deep learning. L'exemple *ci-dessus* à gauche est un système de captation de photos dans la bijouterie d'**Ortery**. L'appareil photo est monté sur un rail cylindrique et les objets sont placés sur un plateau tournant. Il existe un grand nombre de solutions de ce genre pour capter des photos d'objets pour le commerce en ligne. Ils sont souvent utilisés pour permettre une visualisation interactive de l'objet que l'on peut déplacer avec sa souris.

[367] Voir Some startups use fake data to train AI, de Tom Simonite, avril 2018 et Deep learning with synthetic data will democratize the tech industry, de Evan Nisselson, mai 2018.

[368] Voir Deep learning with synthetic data will democratize the tech industry, mai 2018.

[369] Voir 5 tips to overcome machine learning adoption barriers in the enterprise, de Alison DeNisco Rayome, novembre 2017.

Un cas extrême consisterait à entourer l'objet de murs de LEDs alimentés par des générateurs d'images de synthèse pour recréer une grande variété d'éclairages et situations autour de l'objet à photographier. C'est la technique utilisée pour le tournage de la série **The Mandalorian**, mais pas pour alimenter un réseau de neurones[370] (*ci-dessus* à droite)!

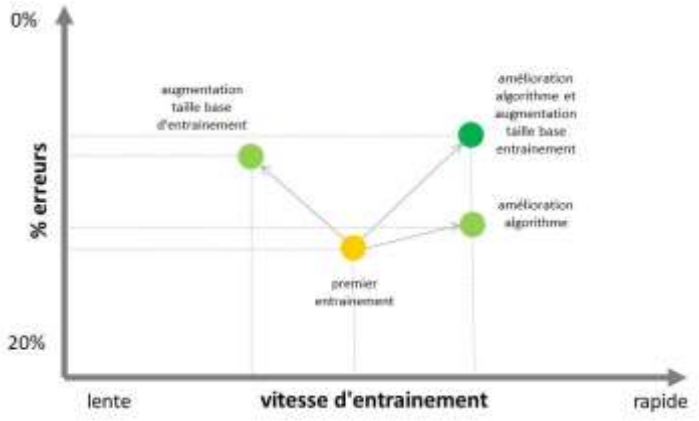
Là encore, les entreprises et les startups devront prendre en compte le règlement européen RGPD dans la collecte et le traitement des données personnelles. Leur portabilité d'un service à l'autre sera l'une des obligations les plus complexes à gérer. Le droit à l'oubli également[371] !

Des scandales émergent lorsque des GAFA ou des startups utilisent des images postées par leurs utilisateurs sans leur demander explicitement leur avis. Cela déclenche quelques tempêtes médiatiques qui se calment généralement rapidement après avoir consommé du budget d'avocats et de lobbyistes pour remettre de l'ordre[372].

Certaines études portant sur un seul type de réseau de neurones montrent qu'une IA avec plus de données est plus efficace qu'une IA avec un meilleur algorithme.

La performance des algorithmes joue cependant un rôle clé dans la qualité des résultats dans le deep learning, et surtout dans leur performance, notamment la rapidité de la phase d'entraînement des modèles. Pour ce qui est de la reconnaissance des images, il faut distinguer le temps d'entraînement et le pourcentage de bonnes reconnaissances. Les progrès des algorithmes visent à améliorer l'une comme l'autre. La taille des jeux de données est en effet critique pour bien entraîner un modèle.

Si l'algorithme utilisé n'est meilleur que dans la vitesse d'entraînement, ce qui est souvent le cas dans des variantes de réseaux de neurones convolutifs, alors, la performance de la reconnaissance ne changera pas lors de l'exécution du modèle entraîné. Par contre, avec plus de données d'entraînement, celui-ci sera plus long. Comme illustré dans le schéma *ci-contre*, il faut à la fois de meilleurs jeux de données et de meilleurs algorithmes pour que l'entraînement soit aussi rapide que possible.



C'est notamment utile pour réduire la consommation énergétique de l'IA. Bref, pour faire de l'IA verte[373] ! Mais il est bon de tenir compte des délais du projet : l'optimisation d'un algorithme pourra prendre beaucoup plus de temps que son alimentation avec un jeu de données plus grand.

---

[370] Voir Why 'The Mandalorian' Uses Virtual Sets Over Green Screen, juin 2020 (6 minutes).

[371] Lorsqu'un réseau de neurones aura été entraîné avec des données personnelles de millions d'utilisateurs, la suppression des données personnelles d'une base de données ne signifiera pas automatiquement qu'elles ont disparu du réseau de neurones entraîné avec. Mais les données utilisées dans l'entraînement sont normalement anonymisées puisqu'elles servent à déteminer des caractéristiques des utilisateurs à partir de paramètres divers (localisation, comportement, usages). Les données ont beau être anonymisées, elles figurent sous la forme d'influence probabiliste du réseau de neurones entraîné. Influence qui est normalement négligeable à l'échelle d'un seul utilisateur. A l'envers, un réseau de neurones bien entraîné peut deviner des caractéristiques cachées d'un client via son approche probabiliste. Ces informations déduites doivent donc probablement être aussi bien protégées que les informations d'origine sur l'utilisateur.

[372] Voir Ever quietly trained facial recognition AI using its photo storage app par Amrita Khalid, mai 2019 et The Ever photo app turned users' private snaps into AI facial recognition fodder par par James Vincent, mai 2019.

[373] C'est un des objectifs du chercheur **Stéphane Canu** de l'INSA Rouen qui planche sur l'optimisation de gros modèles de vision artificielle et de traitement du langage. D'où le projet de recherché collaborative "Deep in France" lancé par différents laboratoires et financé par l'ANR.

Par contre, l'économie de quelques points d'erreurs à la marge est très coûteuse en données d'entraînement et avec une progression exponentielle, sachant que cela dépend aussi de la variabilité des situations à analyser en production.

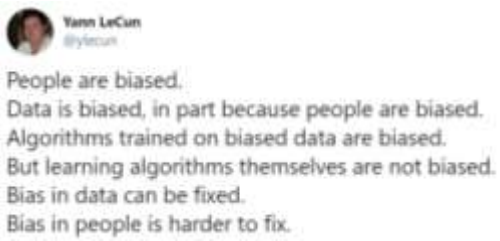
Où ces données sont-elles stockées ? Elles peuvent l'être sur les serveurs de l'entreprise ou dans le cloud et si possible dans un cloud bien privé de l'entreprise. Contrairement à une idée répandue, les services de cloud issus des GAFAMI n'exploitent pas les données des entreprises qui y sont stockées. Seules celles qui proviennent des services grand public (moteurs de recherche, réseaux sociaux, email personnels) peuvent l'être.

Par contre, les données qui circulent sur Internet peuvent être interceptées par certains services de renseignement qui ont installé des sondes sur les points d'accès des grandes liaisons intercontinentales. La DGSE le fait pour les fibres qui arrivent en France et la NSA pour celles qui arrivent aux USA, en général à des fins de renseignement sur le terrorisme mais cela peut déborder sur d'autres besoins !

Biais des données

Le biais des IA est souvent évoqué car il peut affecter les résultats des traitements de machine learning et de deep learning. Mais le biais le plus fort est issu des données qui les alimentent[374]. Et ces biais sont souvent inconscients et pas anticipés par les développeurs et certains data scientists[375].

Comme l'indiquait bien Yann Le Cun en décembre 2019, le deep learning n'est pas biaisé par nature[376]. Il ne fait que reproduire les biais, souvent d'origine humaine, contenus dans les données d'entraînement. Il se bat contre l'idée répandue que ce sont les algorithmes qui sont biaisés[377]. En fait, il faudrait peut-être biaiser les algorithmes pour corriger les biais des données qui les alimentent !



Deux anecdotes l'illustrent parfaitement : chez Facebook, les femmes afro-américaines se sont rendu compte en 2016 que les détecteurs de main dans les distributeurs de savon dans les WC ne fonctionnaient pas avec elles.

Pour ces mêmes personnes, certains systèmes de reconnaissance de visages ne fonctionnent pas mieux. Pourquoi donc ? Dans le premier cas, cela peut être lié au capteur utilisé. Dans le second, c'est une histoire de données d'entraînement qui ont alimenté le système de reconnaissance de visage.

Le point commun : les créateurs de ces systèmes n'avaient pas de personnes de couleur dans leurs équipes techniques. D'où un biais dans le matériel, dans les logiciels et les données.

métaphore du biais de la mesure

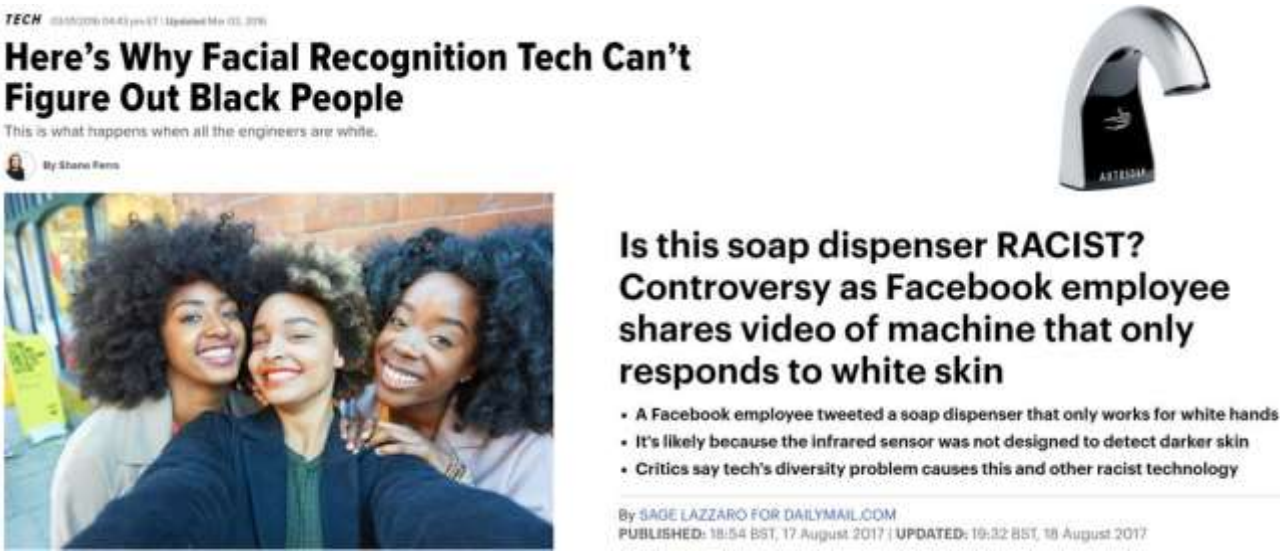
[374] Voir l'ouvrage de référence Weapons of Math Destruction de Cathy O'Neil et cette vidéo d'une heure où elle résume son propos ainsi que son interview Les algorithmes exacerbent les inégalités, novembre 2018.

[375] Voir How Bias Distorts AI (Artificial Intelligence) par Tom Taulli, 2019.

[376] Dans un flux Twitter de décembre 2019 qui avait déclenché une violente polémique, à la suite de laquelle Yann Le Cun s'était temporairement retiré de Twitter.

[377] Voir Yann LeCun supprime son compte Twitter suite à une discussion sur la partialité de l'IA avec ses pairs, les systèmes d'IA sont biaisés s'ils sont préformés avec des données biaisées, dit-il par Bill Fassinou, juillet 2020.

Dans la reconnaissance de visages, le biais vient de ce que les données comprenaient une proportion trop faible de visages de couleur. Et il semble que cela persiste dans les systèmes d'aujourd'hui selon une étude du MIT[378].



TECH

Here's Why Facial Recognition Tech Can't Figure Out Black People

This is what happens when all the engineers are white.

By Shane Ferro

Is this soap dispenser RACIST? Controversy as Facebook employee shares video of machine that only responds to white skin

- A Facebook employee tweeted a soap dispenser that only works for white hands
- It's likely because the infrared sensor was not designed to detect darker skin
- Critics say tech's diversity problem causes this and other racist technology

By SAGE LAZZARO FOR DAILYMAIL.COM
PUBLISHED: 18:54 BST, 17 August 2017 | UPDATED: 19:32 BST, 18 August 2017

Des méthodes de débiaisage des données d'entraînement existent, elles aussi créés par le MIT. Elles consistent à modifier dynamiquement le poids des différents types de visages dans les données d'entraînement[379]. Chez DeepMind, des méthodes similaires évitent les biais de genre dans les traductions, proposant aussi bien le genre masculin que féminin et autres variétés[380].

Pour les images, une technique consiste à faire disparaître des parties d'images et de vérifier quelles parties sont utilisées par le modèle pour reconnaître l'objet. Cette technique peut aussi être employée pour débiaiser une sélection de CV en cachant par exemple le prénom, l'adresse et identifier les principaux éléments du CV qui serviront à sélectionner le candidat.

Une IA doit donc être alimentée par des jeux de données d'entraînement qui sont les plus représentatives des usages à couvrir, et notamment en termes de diversité d'utilisateurs. Cela demande de l'empathie et cela exige pour les créateurs de ces solutions de sortir de leurs cadres de vie et de pensée habituels[381]. En termes statistiques, cela veut dire que les données doivent avoir un fort écart type et une distribution similaire à celle des usages visés[382]. Les données d'entraînement d'IA qui portent sur le fonctionnement de machines doivent répondre aux mêmes exigences.

La diversité de l'équipe des data scientists et des développeurs de modèles d'IA (genre, culture, nationalité…) permet aussi de réduire les biais des systèmes.

---

[378] Voir Study finds gender and skin-type bias in commercial artificial-intelligence systems, de Joy Buolamwini dans MIT News, février 2018 (video).

[379] Voir MIT hopes to automatically 'de-bias' face detection AI par Jon Fingas, janvier 2019.

[380] Voir Google Translate gets rid of some gender biases par Megan Rose Dickey, décembre 2018.

[381] Voir Forget Killer Robots—Bias Is the Real AI Danger de John Giannandrea (Google), octobre 2017,

[382] En juin 2018, IBM annonçait la diffusion à venir d'une base contenant un million de visages en open data, avec une bonne représentation de la diversité. Elle serait cinq fois plus grande que la plus grande base ouverte actuelle. La base sera complétée d'une base encore mieux labellisée de 36 000 visages représentant de manière identique l'ensemble des ethnies, ages et genres. Voir https://www.ibm.com/blogs/research/2018/06/ai-facial-analytics/.

Bref, pour faire du machine learning, il faut conserver de solides notions de statistiques et probabilités et ne pas se faire embobiner par des promesses douteuses[383]. Il faut aussi une incitation économique à bien faire les choses et à en valider le bon fonctionnement [384]!

En juillet 2018, l'association **ACLU** (American Civil Liberties Union) fit un test de croisement entre les 535 membres du Congrès US et 25 000 visages d'une base de criminels en open data avec le SDK Rekognition d'Amazon[385]. Le système reconnu 28 criminels de la base dans les 535 élus ! Amazon s'est défendu en mettant en avant la configuration utilisée qui tolérait 20% d'erreurs alors qu'ils recommandent à la police d'utiliser un seuil de 5%. On est ici dans un cas de biais de configuration d'algorithme plus que de données.

Ainsi, si on entraîne une IA à reconnaître le bruit de moteurs en panne, il faut disposer d'une base d'entraînement de bruits de moteurs représentative des divers types de pannes qui peuvent affecter les-dits moteurs. Sinon, certaines pannes ne seront pas détectées en amont de leur apparition.

Un autre biais peut affecter le machine learning lorsque celui-ci est affecté à la réalisation de prévisions : le biais des « **données-rétroviseur** ». En effet, les données du passé ne correspondent pas à l'avenir ni à un présent acceptable. Vous en êtes les témoins au quotidien lorsque vous planifiez un voyage dans un pays ou cherchez un produit donné, puis faites votre achat. Dans les jours et mois qui suivent, vous pourrez alors être bombardés de reciblage publicitaire sur ces mêmes pays et produits alors que vous n'en avez plus besoin.

Tous les achats ne relèvent pas de « repeat business » ! Si des données du présent reflètent des disparités et des inégalités, et qu'elles sont associées à des systèmes de recommandation, les IA ne feront que perpétuer ces disparités et inégalités.

On en a eu une autre belle démonstration avec cette prévision des résultats de la Coupe du Monde FIFA 2018[386] réalisée en juin 2018 par Goldman Sachs. Elle anticipait une victoire du Brésil contre l'Allemagne en finale et une défaite de la France contre le Brésil en demi-finale.

Tout en indiquant que, par essence, le football était un sport difficile à prédire. Alors, pourquoi donc faire des prévisions ? Qui se sont bien évidemment royalement plantées ! On pourrait même dire qu'il était statistiquement hautement probable que ces prévisions soient à côté de la plaque tant la théorie du chaos règne dans une telle compétition[387].

Autre exemple connu, le cas de **Google Flu** qui tentait de prédire les épidémies de grippe en se basant sur les recherches portant sur le sujet. Cela fonctionnait bien en moyenne mais l'outil a loupé l'épidémie de 2013, notamment aux USA.

---

[383] L'illustration de potirons illustre de manière imagée ce besoin de diversité dans les données. Elle est extraite d'une présentation de Nikos Paragios lors de la conférence FranceIsAI d'octobre 2018 à Station F, à Paris. Voir aussi Using Artificial Intelligence To Customize Content Without Bias par Jennifer Kite-Powell, 2019.

[384] Voir Why Is AI And Machine Learning So Biased? The Answer Is Simple Economics par Kalev Leetaru, 2019 qui décrit le manque d'incitations économiques à réduire les biais des IA.

[385] Voir Amazon's Facial Recognition Matched Congress Members to Criminals, juillet 2018.

[386] Voir Goldman Sachs used AI to simulate 1 million possible World Cup outcomes — and arrived at a clear winner, juin 2018.

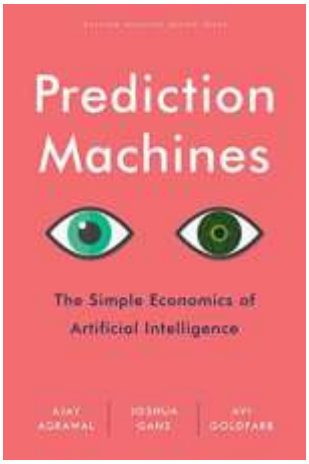
[387] Voir Coupe du Monde : le big data s'est encore spectaculairement raté dans ses prévisions, de Sylvain Rolland dans La Tribune, juillet 2018 qui fait état d'autres prévisions qui étaient tout autant à côté de la plaque que celle de Goldman Sachs.

Même si l'outil n'utilisait probablement pas de deep learning, il est probable que si cela avait été le cas, il lui aurait été difficile de prédire l'épidémie[388]. Depuis, Google Flu a été abandonné.

Une mésaventure du même type a été découverte chez **Amazon** dont le système d'assistance au recrutement avait tendance à défavoriser les femmes candidates car les données d'entraînement comportaient très peu de femmes, comme c'est le cas dans de nombreux métiers techniques. La sous-représentation statistique des femmes dans le modèle avait donc tendance à diminuer leurs chances d'être retenues comme candidates. Cela montre que si l'on souhaite que les IA ne perpétuent pas des situations existantes insatisfaisantes, il faut modifier les représentations statistiques des données d'entraînement pour aller vers un équilibre meilleur ou plus équitable selon les critères du moment[389]. Là encore, il s'agit d'éviter l'effet rétroviseur des données du passé.

Une entreprise qui dépendrait trop des données du passé pour prédire le futur est l'analogue d'un conducteur de véhicule dont le rétroviseur aurait la taille du pare brise et conduirait en pleine nuit ! Le platane serait difficile à éviter dans ces conditions. C'est bien évidemment métaphorique mais explique pourquoi il faut prendre des précautions lorsque l'on fait reposer une décision sur les données du passé. Ce biais du rétroviseur est celui qui conduit les IA à être sexistes ou racistes, parce que les contenus utilisés pour l'entraîner reflètent le monde réel… qui l'est [390]! Les données du réel sont bien plus biaisées que ne le sont les algorithmes qui les exploitent, ce qui peut d'ailleurs conduire à confondre les deux lorsque l'on dénonce les biais sexistes de l'IA[391].

Finalement, le défaut des IAs est de trop refléter le monde existant et pas assez le monde souhaitable ! Et ce monde souhaitable ne peut être obtenu qu'en injectant un biais explicite dans le choix des données, pas une sorte de biais subi sans avoir réfléchi. C'est une question autant politique que technique. Il en va de même pour des applications éminentes de l'IA que sont les agents vocaux, assez souvent féminins (Alexa, Cortana…), les robots sexuels ou les armes robots. Ici, les données n'y sont pour rien. Il s'agit de choix sociologiques, marketing et politiques de leurs créateurs[392].



Dans l'ouvrage **Prediction Machines**, Ajay Agrawal, Joshua Gans et Avi Goldfarb[393] indiquent ainsi que les entreprises devront faire preuve de discernement pour allouer des tâches de prévisions aux IAs ou aux humains. Bien entendu !

Les biais d'incomplétude interviennent couramment dans le traitement du langage lorsque des expressions sont interprétées mot à mot au lieu de l'être comme idiomatismes.

---

[388] Voir Deep Learning: Diminishing Returns? de Bernard Murphy, juillet 2018.

[389] Voir Pourquoi il faut défendre Amazon et son algorithme de recrutement d'Aurélie Jean, dans le Point, octobre 2018. Voir aussi cette histoire d'un logiciel de recrutement à base de règles d'une faculté de médecine aux USA dans les années 1970/1980 : Untold History Of AI: Algorithmic Bias Was Born In The 1980s par Oscar Schwartz, avril 2019.

[390] Voir Les intelligences artificielles sont-elles sexistes ? Des spécialistes nous répondent de Mathilde Saliou, février 2018.

[391] Comme dans MACHORITHMES : les algorithmes sont-ils sexistes ?, par Axelle Lemaire, décembre 2020. Dans la majorité des cas, les biais sexistes, s'ils peuvent provenir de concepteurs masculins, proviennent surtout de la mécompréhension de l'impact statistique de données du réel qui reflètent les biais de la société. Les algorithmes sont souvent neutres de ce point de vue-là. La solution consiste donc soit à biaiser statistiquement dans le bon sens les données d'entraînement (avec plus de femmes par exemple) ou à biaiser statistiquement les algorithmes qui les exploitent. Le premier est plus facile à faire que le second de manière générique.

[392] Voir AI has a gender problem. Here's what to do about it, de Samir Saran et Madhulika Srikumar, avril 2018.

[393] Cités dans The Economics Of Artificial Intelligence - How Cheaper Predictions Will Change The World dans Forbes en juillet 2018.

Une manière de visualiser cela est d'imaginer ce qu'un réseau de neurones génératif créerait comme image pour les expressions familières « mettre les pieds dans le plat » et « parler la langue de bois », comme vu *ci-contre*. Les métaphores imagées ne sont pas encore le fort des IA génératives !

D'autres surprenant biais peuvent provenir des données d'entraînement utilisées et de leur manque de diversité. Ainsi, des objets peuvent être reconnus dans une image en fonction seulement de leur contexte et pas de leurs caractéristiques. C'est le cas avec cet exemple de huski (à gauche) qui est confondu par un réseau de neurones convolutionnel avec un loup car les éléments de l'image qui le discriminent sont ceux du fond de neige. Ce genre de biais est une forme « d'overfitting ».

Cela vient du fait que la base d'apprentissage comprenait quasiment exclusivement des animaux pris sur un fond neigeux.

La manière de s'en rendre compte est d'utiliser un réseau de neurone de segmentation qui va générer une « heat map » des zones de l'image qui ont servi à reconnaître l'animal. Et là, on se peut se rendre compte de l'erreur.

L'autre biais humain classique est de confondre corrélation et causalité. Les exemples abondent dans ce sens comme cette vague corrélation entre la consommation de chocolat dans un pays et son nombre de prix Nobel par habitant.

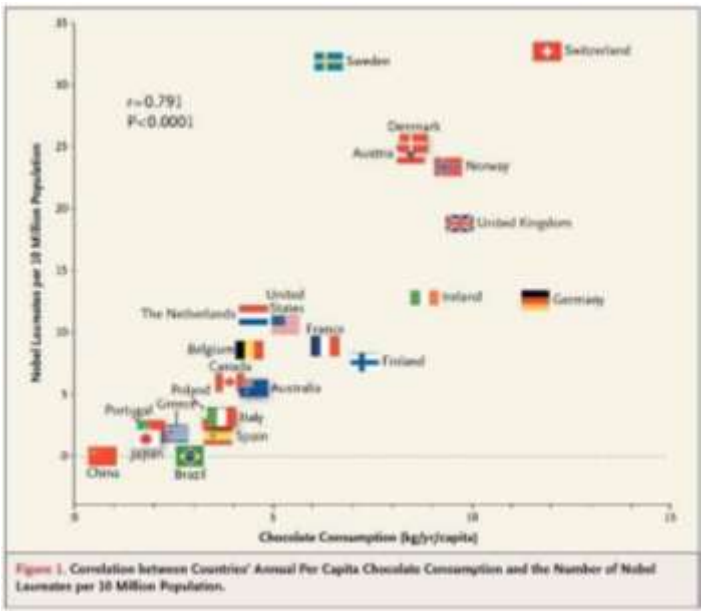
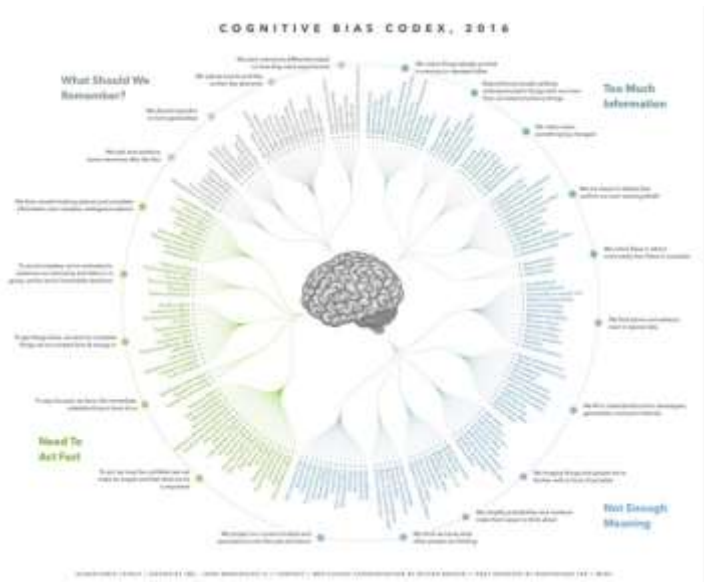


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

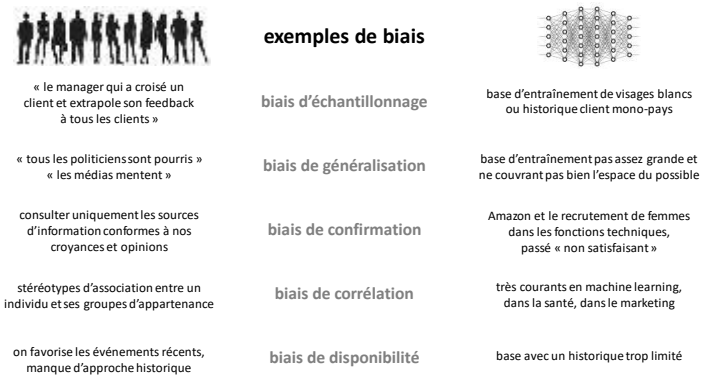
Dans un tel cas, il existe au moins une demi-dizaine d'autres critères qui influent ces deux paramètres. Le machine learning entraîné uniquement avec des données de consommation de chocolat et de prix Nobel n'y verra que du feu. Par contre, une analyse multi-paramètres sera peut-être pertinente. Ici, c'est le choix de la structure des données analysées qui influera le résultat.

Dans d'autres cas de figure, une corrélation est explicable, comme celle-ci, *ci-contre*, portant sur le lien entre le port du masque et la diffusion du covid-19 au début de la pandémie en mars 2020. Cela étant, les masques protègent bien du virus, mais ce n'est pas la seule donnée à prendre en compte. Il y a aussi la distanciation, la densité de la population et des transports en commun, la température, le groupe sanguin, et plein d'autres paramètres que l'on ne connaît pas forcément encore.

Mais dénoncer les biais de l'IA revient aussi à oublier que les humains ont aussi des biais qui sont bien pires en général. La cartographie ci-contre les classifie et les inventorie[394]. Cela en fait beaucoup ! Les humains utilisent couramment un mix d'IA connexionniste (statistique) et symbolique (raisonnement). On observe le monde, on en déduit rapidement des généralités souvent abusives, et ces généralités deviennent des règles qui régissent notre comportement. On est régulièrement victimes de biais divers comme le biais d'échantillonnage lié à un échantillon insuffisant et une extrapolation abusive.



Les biais de généralisation, de confirmation, les biais de corrélation (vus *ci-dessus*) et les biais de disponibilité ont tous trait aux insuffisances des données servant à « nous » entraîner. On peut ainsi faire une petite correspondance simple entre les biais humains et ceux des IA.



N'est pas forcément le plus biaisé qui croit. En fait, nous sommes manipulables par peu d'information[395].

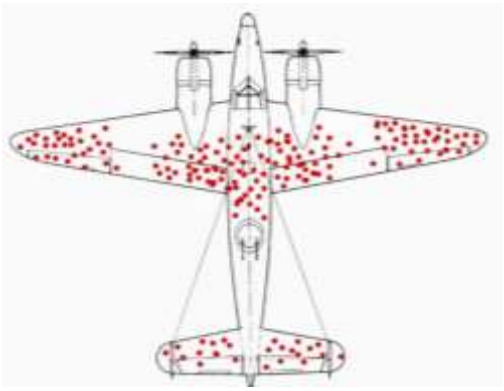
Les cartographies de biais ne manquent pas. En voici donc une autre, bien nourrie et fournie d'exemples concrets[396].

[394] Voir Cognitive bias cheat sheet - Better Humans - Medium par Buster Benson, septembre 2916 et la version en français.

[395] Voir IA : Nous sommes beaucoup plus prévisibles que nous le pensons ! par Stéphane Mallard, mars 2019.

[396] Source : 50 Cognitive Biases in the Modern World par Marcus Lu, février 2020.

Il existe un biais bien connu, celui du survivant.



Il consiste à analyser les caractéristiques des survivants d'une épreuve en oubliant d'analyser celles des non survivants. L'illustration la plus évidente est celle de ces bombardiers de la seconde guerre mondiale dont on analysait les parties touchées par la flak allemande pour renforcer le blindage. En fait, ces parties étaient les moins vulnérables. Abraham Wald, un statisticien hongrois émigré aux USA, découvra le point clé : les parties non touchées étaient les plus vulnérables et correspondaient aux avions qui n'étaient pas revenus ! C'est là qu'il fallait renforcer le blindage[397] !

Ce biais du survivant se manifeste inversement dans nombre de présentations de solutions de deep learning ou de réseaux génératifs. Les présentations qui en sont faites mettent souvent en avant les cas où le système fonctionne bien et évitent de décrire le cas où cela ne fonctionne pas bien.

On peut aussi évoquer le fameux effet **Dunning-Kruger** qui veut que les novices surestiment toujours leurs capacités par rapport aux experts. Sa contraposée est le syndrome de l'imposture. La métaphore est connue : les andouilles ne doutent de rien ! Cet effet n'a pas vraiment d'application contrète connue dans l'IA.

Comment donc éviter ces biais, qu'ils viennent des données ou des algorithmes du machine learning ? Des outils d'analyse sont proposés par diverse entreprises et projets open source qui servent à vérifier que les résultats du machine learning ne génèrent pas de biais, par exemple sur des groupes d'individus selon la race, l'âge ou le genre. Nous avons par exemple **IBM AI Fairness 360** (AIF360) lancé en 2018 (vidéo), **Facebook** Fairness Flow lancé en mai 2018, **Pymetrics** Audit AI, **Fairness Measures** (test d'algorithme avec jeux de données variés et indicateurs d'équité), **Fairness Comparison**[398], **Themis-ML** (bibliothèque Python pour Scikit-Learn qui génère des algorithmes de machine learning respectant l'équité[399], issu de la startup **Arena** (2009, USA, $7,5M)), **Aequitas** Bias and Fairness Audit Toolkit (audit de biais) et **Fairtest**[400] (outil de tests).

---

[397] L'histoire est bien racontrée dans Survivorship Bias par David McRaney, 2013. D'ailleurs, sur le schéma, on peut intuiter qu'il ne serait pas inutile de mieux protéger les moteurs ou le poste de pilotage qui sont plus vitaux qu'un bout d'aile.

[398] Voir A comparative study of fairness-enhancing interventions in machine learning par Sorelle A. Friedler et al, février 2018 (22 pages).

[399] Voir Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation par Niels Bantilan, octobre 2017 (8 pages).

[400] Voir FairTest: Discovering Unwarranted Associations in Data-Driven Applications par Florian Tramèr et al, 2017 (16 pages).,

Audit de l'IA

Une fois que l'on a bien assimilé les questions liées au biais des données, à leur sélection, leur préparation et à leur labellisation, puis intégré les enjeux d'éthique de l'IA, il faut pouvoir auditer une IA. Soit a priori, avant sa mise en production, soit après coup, et notamment en cas d'incident. C'est particulièrement important pour les applications critiques comme dans la santé, la sécurité, la justice et dans les moyens de transport autonomes[401].

Il ne faut d'ailleurs pas confondre l'audit de l'IA et l'IA utilisée dans les audits financiers comme avec la solution AI Auditor de **MindBridge** (2015, Canada, $42,3M).

Le champ de l'audit de l'IA est encore jeune.

Voici quelques pistes pour intégrer ce sujet :

- Analyser les **bonnes pratiques** de l'industrie dans les secteurs sensibles et dans votre secteur d'activité.
- Bien documenter les **processus liés aux données** : sélection, préparation et labellisation.
- Auditer la **sécurité d'ensemble** de la chaîne de valeur de l'application de l'entraînement à la mise en production.
- Détecter les **biais potentiels** des données utilisées dans vos applications, voire les biais générés par les capteurs eux-mêmes ou leurs défaillances techniques potentielles.
- Créer des **IA explicables** en fonction de l'état de l'art. Notamment en associant des briques d'IA symbolique et d'IA connexionistes.
- Intégrer les **humains dans la boucle** tant que possible, en particulier lorsque l'IA va générer une information aidant à prendre une décision critique.
- Etudier la **jurisprudence** qui se construit progressivement, comme pour les accidents de véhicules autonomes donc nous reparlerons plus loin.
- Créer des **équipes interdisciplinaires** dans l'entreprise associant maîtrise d'ouvrage, maitrise d'œuvre, développeurs, data scientists, responsables de la sécurité informatique, équipes juridiques.

Capteurs et objets connectés

Les capteurs et objets connectés jouent un rôle clé dans de nombreuses applications d'intelligence artificielle. Les micros et caméras alimentent les systèmes de reconnaissance de la parole et de vision artificielle. Les smartphones et les outils d'accès à Internet en général créent des tombereaux de données sur les comportements des utilisateurs. La ville intelligente (smart city) et les véhicules autonomes sont aussi alimentés par moult capteurs en tout genre.

L'un des moyens de se rapprocher et même de dépasser l'homme est de multiplier les capteurs sensoriels. La principale différence entre l'homme et la machine réside dans la portée de leurs capteurs. Pour l'homme, la portée est immédiate et ne concerne que ses alentours. Pour les machines, elle peut être distante et globale.

On voit autour de soi, on sent la température, on peut toucher, etc. Les machines peuvent capter des données environnementales à très grande échelle. C'est l'avantage des réseaux d'objets connectés à grande échelle, comme dans les "smart cities". Et les volumes de données générés par les objets connectés sont de plus en plus importants, créant à la fois un défi technologique et une opportunité pour leur exploitation.

---

[401] Voir High-Stakes AI Decisions Need to Be Automatically Audited par Oren Etzioni, Wired, juillet 2019.

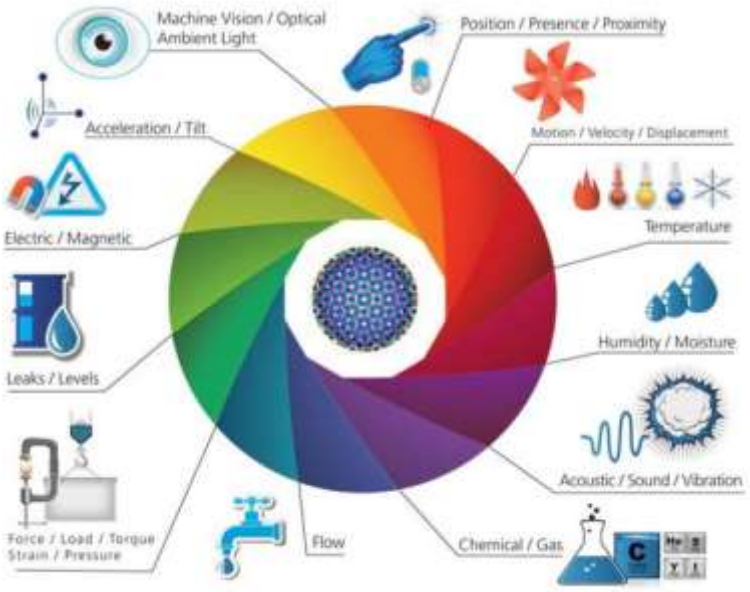
Le cerveau a une caractéristique méconnue : il ne comprend pas de cellules sensorielles. Cela explique pourquoi on peut faire de la chirurgie à cerveau ouvert sur quelqu'un d'éveillé. La douleur n'est perceptible qu'à la périphérie du cerveau. D'ailleurs, lorsque l'on a une migraine, c'est en général lié à une douleur périphérique au cerveau, qui ne provient pas de l'intérieur. L'ordinateur est dans le même cas : il n'a pas de capteurs sensoriels en propre. Il ne ressent rien s'il n'est pas connecté à l'extérieur. Une IA sans capteurs ni données ne peut pas faire grand chose.

Cette différence peut même se faire sentir à une échelle limitée comme dans le cas des véhicules à conduite assistée ou automatique qui reposent sur une myriade de capteurs : ultrasons, infrarouges, vidéo et laser / LIDAR, le tout fonctionnant à 360°. Ces capteurs fournissent aux ordinateurs de bord une information exploitable qui va largement au-delà de ce que le conducteur peut percevoir, surtout dans la mesure où les données de plusieurs capteurs sont combinées (« sensor fusion »).

C'est l'une des raisons pour lesquelles les véhicules autonomes sont à terme très prometteurs et plus sécurisés. Ces techniques sont déjà meilleures que les sens humains, surtout en termes de temps de réponse, de vision à 360° et de capacité d'anticipation des mouvements sur la chaussée (piétons, vélos, autres véhicules). A contrario, la finesse de la vue humaine n'est pas encore égalée par la vision artificielle de fait de ses contraintes actuelles.

En effet, les réseaux de neurones convolutifs utilisent pour l'instant des images sources à basse résolution pour tenir compte des contraintes matérielles actuelles, en particulier de mémoire vive. Ils sont de plus rares à fonctionner en 3D avec une vision stéréoscopique[402].

Le marché des capteurs a connu un fort développement depuis la fin des années 2000 grâce à l'émergence du marché des smartphones, alimenté par l'iPhone et les smartphones Android. Il s'en vend actuellement environ 1,5 milliards d'unités par an et ils sont renouvelés à peu près tous les deux ans par les consommateurs. N'importe quel smartphone comprend au minimum une douzaine de capteurs : deux à six caméras, un à deux micros, un accéléromètre, un gyroscope, un GPS, un capteur de lumière, un capteur de proximité, un capteur d'empreintes digitales et des capteurs radio Bluetooth / Wifi / 2G / 3G / 4G / 5G.

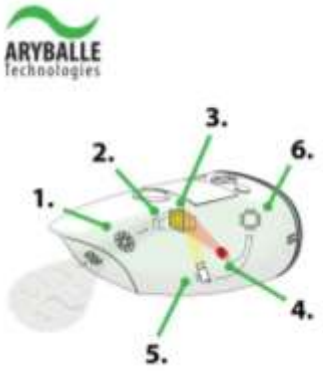


Cela a eu comme conséquence d'accélérer la miniaturisation et la baisse du prix de tous ces capteurs. Les innovations dans le secteur des capteurs se poursuivent à un bon rythme et permettent de créer des perceptions extra-sensorielles par rapport aux capacités humaines.

Chacun de ces capteurs va générer des données exploitables par des systèmes de machine learning pour comparer le signal acquis avec des bases de données de signaux déjà associés à des matières déjà détectées.

[402] Voir un exemple dans Usings CNNs to Estimate Depth from Stereo Imagery de Tyler S. Jordan er Skanda Shridhar, 2014 (6 pages) ainsi qu'avec 3D Facial Expression Reconstruction using Cascaded Regression 2018 (8 pages).

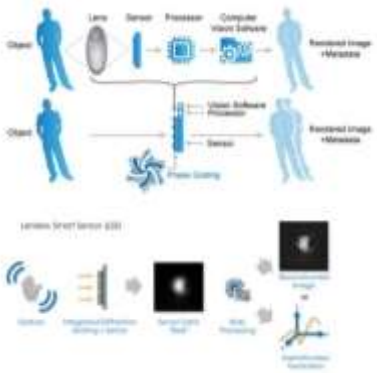
Nous en avons deux exemples avec les spectrographes infrarouges comme ceux de l'israélien **Scio**, intégrés dans une balance de **Terraillon** lancée en 2017 ou un smartphone de Changhong lancé au CES 2017, mais qui ne semble cependant pas avoir percé ni été renouvelé les années suivantes, avec le détecteur de gaz Neose du français **Aryballe** ou encore avec le détecteur de pollution aérienne d'une autre startup françaiss, **Plume Labs**.



Les plateformes de gestion de maisons connectées tirent aussi parti de nombre de capteurs d'ambiance pour optimiser le confort. Ils jouent sur l'intégration de données d'origine disparate : la température extérieure et intérieure, l'humidité, la luminosité ainsi que les déplacements des utilisateurs, captés avec leur smartphone. Cela permet par exemple d'anticiper la température du logement en anticipation du retour au domicile de ses occupants.

Cette orchestration passe de plus en plus souvent par de l'apprentissage profond pour identifier les comportements des utilisateurs et adapter les réponses du système.

L'innovation dans les capteurs photo et vidéos est également incessante, ne serait-ce que par la miniaturisation de ceux qui équipent les smartphones et sont maintenant dotés de vision en 3D. L'américain **Rambus** planche de son côté sur un capteur photo qui n'a pas besoin d'optique ! Les capteurs de vibrations et les microphones ont des applications industrielles insoupçonnées et révélées par l'IA : la détection d'anomalies.

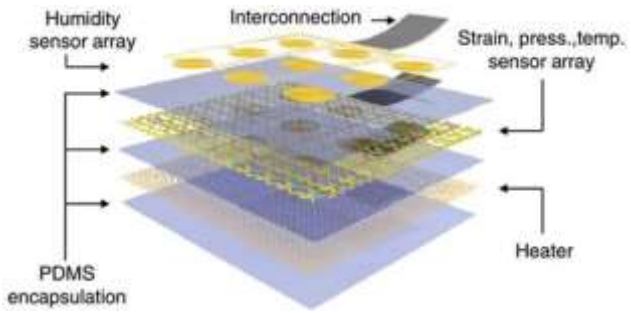


Ainsi, des capteurs placés dans des véhicules ou des machines industrielles génèrent un signal qui est analysé par des systèmes de deep learning capables d'identifier et caractériser les anomalies. Ainsi, **Cartesiam** (2016, France, 2,5M€) a créé NanoEdge AI, une solution logicielle intégrée pour capteurs de vibration intégrant un réseau de neurones servant à détecter les vibrations anormales[403]. Elle est embarquable dans des micro-contrôleurs basse-consommation, d'où un partenariat avec STMicroelectronics, fournisseur des micro-contrôleurs STM32 mais aussi avec NXP et Renesas. La solution est complétée par NanoEdge AI Studio pour le développement des solutions et est exploitée chez des industriels tels que Schneider Electric, Crouzet, Naval Group, Bosch et Altyor.

La startup française a un bureau à New York depuis 2019. Ils ont aussi créé un capteur générique, Bob Assistant, fabriqué par Eolane.

[403] Voir Eolane et Cartesiam marient capteur et intelligence artificielle pour la maintenance prédictive des équipements industriels, avril 2018 et NanoEdge AI: Their First Machine Learning Application on the STM32G4 Series Blew Our Minds, STMicroelectronics, juin 2019. En pratique, Cartesiam propose un framework de machine learning propriéraire embarqué dans les micro-contrôleurs. Ils récupèrent les eux de données des clients, créent un modèle qui est ensuite exécuté dans leur framework. Leur système fonctionne avec apprentissage supervisé et par renforcement, en local.

Les capteurs de proximité intégrables à des machines comme les robots progressent même dans leur biomimétisme. Des prototypes de peau artificielle sensible existent déjà en laboratoire, comme en Corée du Sud (*ci-contre*, **source dans Nature**).



La mécanique humaine la plus difficile à reproduire est celle des muscles.

Ceux-ci restent une mécanique extraordinaire, économe en énergie, fluide dans le fonctionnement, que les moteurs rotatifs électriques des robots ont bien du mal à imiter. Des travaux de recherche sporadiques sont menés dans ce domaine mais n'aboutissent pas encore.

Les **capteurs biométriques** sont de plus en plus courants : dans les bracelets type FitBit, dans les montres connectées avec leurs capteurs infrarouges détectant le pouls et l'oxygénation du sang (comme la Watch d'Apple) et dans les capteurs d'électroencéphalogrammes (EEG). Ces derniers permettent à l'homme de contrôler un membre artificiel robotisé, une application pouvant restaurer des fonctions mécaniques de personnes handicapées, voire démultiplier la force de personnes valides, dans les exosquelettes dédiés aux applications militaires ou dans le BTP. L'homme peut ainsi piloter la machine car la périphérie du cortex cérébral contient les zones où nous commandons nos actions musculaires.

Les caméras dans le visible et l'infrarouge couplées à d'éventuels autres capteurs permettent de détecter l'état psychologique de personnes à distance, comme leur niveau d'intérêt dans une conférence ! C'est un cas d'usage de la société française **datakalab** qui propose cela dans les conférences et même pour les utilisateurs d'Internet, en captant les émotions visuellement et via un bracelet connecté (porté volontairement par des utilisateurs cobayes).

Un dernier point clé à noter : les données issues des capteurs expliquent en partie les prévisions d'accumulation de données faramineuses des analystes. Mais on n'est pas obligé de les stocker de manière centralisée. C'est là que l'IA embarquée dans les objets peut intervenir, pour interpréter les données captées, comme des images, et n'envoyer via une liaison distante que des alertes en fonction des objets et situations détectées et des images seulement si nécessaire après survenue d'un incident.

Infrastructure de l'IA

Les avancées technologiques dans le numérique ont toujours associé le matériel et le logiciel. Comme l'illustre l'histoire d'Ada Lovelace, celle des réseaux de neurones ou du calcul quantique, les innovations algorithmiques ont souvent précédé celles du matériel. Mais l'avènement de matériels à la hauteur des logiciels a toujours été un facteur déclenchant du développement des usages.

Les différentes briques d'infrastructure jouent donc naturellement un rôle critique pour le déploiement de solutions exploitant toute forme d'intelligence artificielle. Cela concerne non seulement les processeurs mais aussi le stockage et les télécommunications. Le rôle des composants a été mis en valeur en 2012 lorsque les GPU ont permis d'accélérer significativement l'entraînement des réseaux de neurones du deep learning, au début, pour la reconnaissance d'images. Depuis, l'industrie des semiconducteurs s'est mobilisée dans la course à l'IA au point de générer de nombreuses catégories de processeurs spécialisés à en perdre son latin et plus de 200 startups actives[404]. Cette explosion cambrienne va probablement donner lieu à un élagage et une consolidation des acteurs lorsque le marché va se sédimenter.

Nous allons ici creuser le rôle et le fonctionnement de ces divers processeurs et de ce qui les accompagne côté mémoire et stockage. Leurs évolutions respectives contribuent aussi à améliorer la qualité et la performance des solutions d'intelligence artificielle.

L'un des outils clés de l'IA sont les serveurs d'entraînement du deep learning. Si celui-ci donne de très bons résultats, comme dans la reconnaissance d'images, il est très consommateur de ressources dans sa phase d'entraînement. Il faut des millions de fois plus de puissance machine pour entraîner un modèle de reconnaissance d'images que pour l'exécuter ensuite dans un smartphone. Ca tombe bien puisqu'un processeur de smartphone doit avoir une faible consommation énergétique.

Cela explique pourquoi, par exemple, les GPU et autres TPU (Tensor Processing Units) ont une capacité de calcul d'environ 100 TFLOPS[405] tandis que les briques neuronales des derniers processeurs de smartphones (Kirin de Huawei, A1x Bionic d'Apple et Qualcomm Snapdragon) se contentent de 2 à 15 TOPS[406]. Et encore, l'entraînement des plus gros réseaux de neurones réclame la mise en batterie de centaines si ce n'est de milliers de serveurs utilisant des GPGPU[407] et TPU[408].

Nous allons aussi aborder les éléments clés de la mise en œuvre de l'IA dans un système d'information : la mémoire, le stockage, les data lakes, les infrastructures du cloud, les outils d'administration à base d'IA (IAops) ainsi que la RPA (Robotic Process Automation) et les outils de la cybersécurité. Nous traiterons également les questions relatives au bilan énergétique des solutions de l'IA.

---

[404] Voir le webinar de deux heures Master class sur le hardware de l'IA organisé par le Hub France IA avec l'auteur, Olivier Ezratty, 18 juin 2020 qui dresse un panorama en deux heures des processeurs de l'IA.

[405] En clair… TFLOPS = tera-floating point operations per seconds. Tera = un millier de milliard. Donc 100 TFLOPS= 100 000 milliards d'opérations en nombres flottants par seconde.

[406] Et petite nuance, la performance des chipsets pour serveurs est souvent indiquée en Flops (opérations en nombres flottants) tandis que celle des composants pour objets connectés de toute sorte est souvent exprimée en Ops (opérations en nombre entiers). Et pour cause : le calcul en nombres entiers consomme moins d'énergie qu'en nombres flottants et on doit privilégier la basse consommation d'énergie dans les mobiles. Résultat : on entraîne des réseaux de neurones en nombres flottants sur serveur et on les exécute en nombre entiers sur mobiles.

[407] General Purpose GPU, les GPU comme ceux de Nvidia qui ont des usages variés associant l'IA et le calcul scientifique.

[408] Tensor Processing Unit, processeurs spécialisés pour le deep learning qui sont optimisés pour faire du calcul vectoriel et matriciel avec des « tenseurs ».

Processeurs

Les applications de l'intelligence artificielles sont de grandes consommatrices de puissance machine. Nous l'avons entrevu en découvrant les principaux algorithmes de réseaux de neurones et du deep learning. Ils manipulent de gros volumes de données.

Un système de reconnaissance d'images utilise un réseau de neurones qui exploite un à plusieurs milliards de paramètres à ajuster pendant son entraînement et à exploiter lors de son exécution. Cela fait une très grande quantité de calculs !

Peut-on compter sur la loi de Moore pour obtenir cette puissance de calcul ? Nous allons voir que oui, mais de manière détournée. Cette loi est la pierre angulaire de nombreuses prédictions technologiques. Présentée comme immuable et quasi-éternelle, cette loi empirique élaborée en 1965 par Gordon Moore anticipait que la densité des transistors dans les processeurs allait doubler tous les 18 à 24 mois selon les versions. Cette loi a été vérifiée empiriquement et la conséquence d'un mix de progrès technologiques et d'enjeux économiques, liés en particulier à la croissance d'Intel.

La loi de Moore est aussi déclinée avec des variantes dans la vitesse des réseaux, la capacité de stockage, le coût d'une cellule solaire photovoltaïque ou celui du séquençage d'un génome humain. Une progression n'en entraîne pas forcément une autre. Le coût peut baisser mais pas la performance brute, comme pour les panneaux solaires photovoltaïques. On peut donc facilement jouer avec les chiffres et passer d'un paramètre à l'autre pour générer cette impression d'exponentielle immuable. La loi de Moore est aussi appliquée de manière naïve pour prédire l'avènement de la fameuse singularité qui verrait l'intelligence artificielle dépasser l'intelligence de l'Homme.

Pour ce qui est de l'IA, les ingénieurs font feu de tout bois pour augmenter la puissance de calcul des machines. Cela passe par l'habituelle course à la densité dans les processeurs CMOS qui atteignent aujourd'hui une densité de 5 nm avec des roadmaps descendant jusqu'à 1 nm, avec des technologies alternatives comme la photonique ou le calcul quantique, puis avec des architectures différentes de processeurs optimisées pour l'entraînement et l'exécution de réseaux de neurones et enfin, avec des technologies diverses qui rapprochent la mémoire du calcul, comme les memristors ou les réseaux de neurones à impulsions. Des processeurs apparaîssent qui sont adaptés à des types particuliers de réseaux de neurones, surtout pour le traitement de l'image et du langage voire du bruit[409]. C'est un phénomène qui n'est pas nouveau. Dans nos ordinateurs et mobiles, les processeurs que l'on appelle des « systems on chip » contiennent déjà des modules spécialisés et performants comme ceux qui traitent l'encodage et le décodage de données audio et vidéo.

Les processeurs de l'IA qui s'inspirent du fonctionnement des neurones et du cerveau sont qualifiés de neuromorphiques. Leur définition est à géométrie variable. En approche extensive, on peut intégrer les processeurs multicoeurs qui s'inspirent des structures corticales du cerveau, les processeurs à base de tenseurs qui réalisent des opérations matricielles permettant de gérer des réseaux convolutifs et du traitement du langage, les processeurs à neurones à impulsion qui s'inspirent des liaisons entre neurones biologiques et les processeurs à memristors qui s'inspirent aussi du fonctionnement des neurones. Chez certains spécialistes, seules les deux dernières catégories sont « neuromorphiques ».

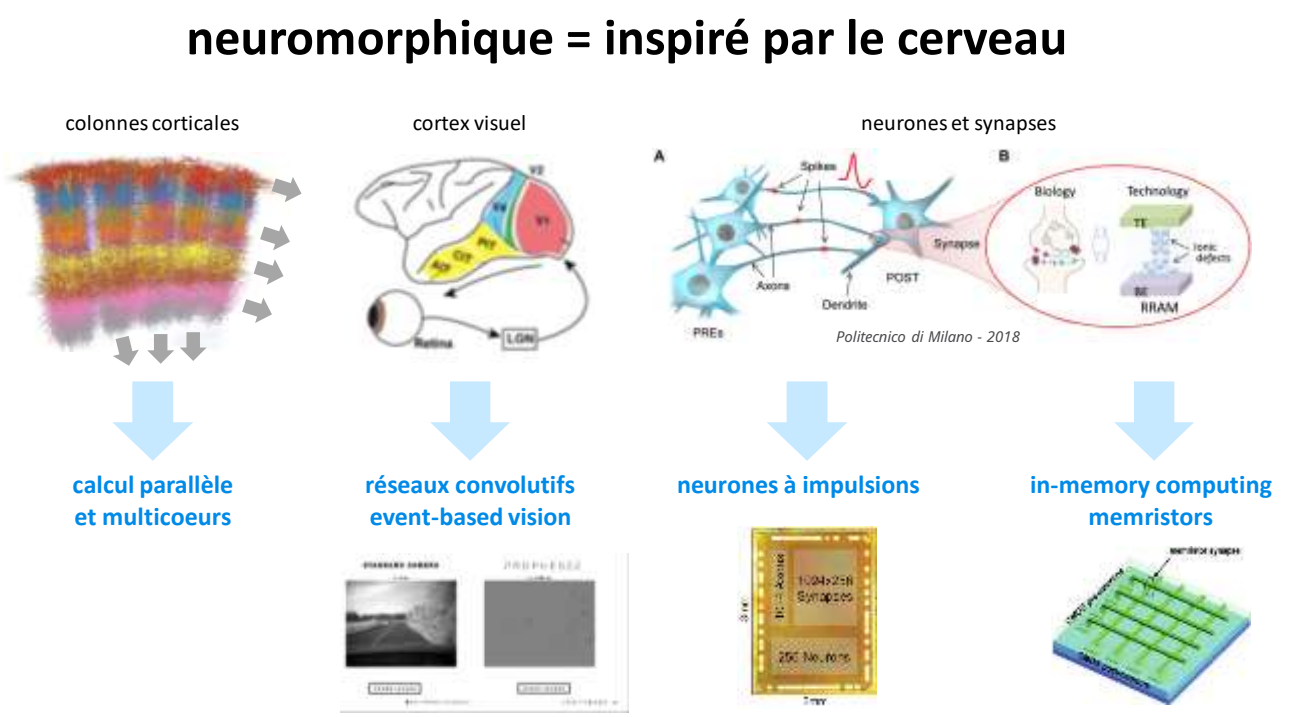
Diverses études récentes montrent que les besoins en calcul dans l'IA augmentent bien plus vite que les effets de la loi de Moore. La principale source de puissance incrémentale provient de la parallélisation des traitements sur de nombreux serveurs, en tout cas côté entraînement, et par un rallongement de leur durée[410].

---

[409] Voir A new golden age for architecture : domain specific hardware/software co-design, enhanced security, open instruction set and agile chip development, juin 2018 (55 slides) qui évoque la notion de DSA, ou Domain Specific Architecture.

[410] Voir AI and Efficiency par OpenAI, 2020 qui fait référence à Measuring the Algorithmic Efficiency of Neural Networks par Danny Hernandez et al, 2020 (20 pages) ainsi que The Computational Limits of Deep Learning par Neil Thompson et al, 2020 (46 pages).

Enfin, signalons que l'IA se sert elle-même, pour la conception à conçevoir des chipsets avec des techniques élaborées d'organisation du layout des transistors et fonctions. C'est une évolution naturelle des outils de conception de chipsets[411].



Course à la densité

Il fut un temps où tous les paramètres des chipsets CMOS évoluaient selon une loi exponentielle ou logarithmique selon les cas : leur densité, le nombre de transistors par chipsets, la vitesse d'horloge ou le prix par transistor. En pratique, la performance dépend de la vitesse d'horloge et de la capacité à paralléliser des traitements.

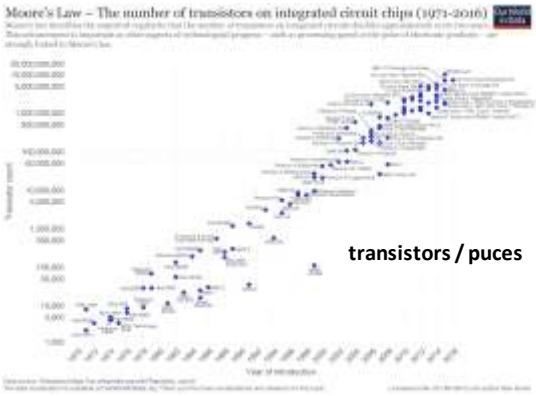
La densité des chipsets et leur nombre de transistors continue d'augmenter régulièrement (*ci-dessous à gauche*) mais pas la fréquence d'horloge qui est au taquet aussi bien sur les laptops, desktops et serveurs (*ci-dessous à droite*). Pourquoi donc ? Pour ne pas faire fondre les chipsets et sur les mobiles, pour limiter la consommation d'énergie ! On plafonne ainsi à 2 GHz et quelques sur mobile et à 4 GHz sur desktops et serveurs, sauf à utiliser des systèmes de refroidissement liquides coûteux. D'où le choix des architectures multicœurs pour paralléliser les traitements qui sont exploitées des serveurs aux mobiles en passant par les ordinateurs personnels.

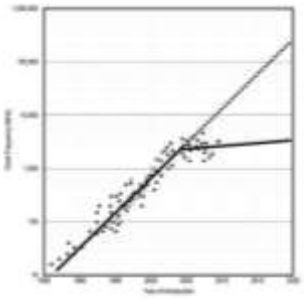
Par ailleurs, il existe une barrière assez méconnue, celle de Landauer, qui définit le niveau minimum d'énergie nécessaire pour changer l'état d'un bit. Elle pourrait être atteinte d'ici 2035. Tout du moins, avec les architectures CMOS, mais pas avec celles de l'informatique quantique que nous verrons plus loin. Cette barrière est cependant contestée par certains physiciens.

La loi de Moore a connu un premier coup d'arrêt vers 2005. A l'époque, les roadmaps d'Intel prévoyaient d'augmenter la fréquence allègrement jusqu'à plus de 20 GHz d'ici 2013. Intel a du rapidement réviser ses prévisions, d'abord à 7 GHz, puis, dans la pratique, est resté en-dessous de la barrière des 5 GHz pour les processeurs de PC et serveurs.

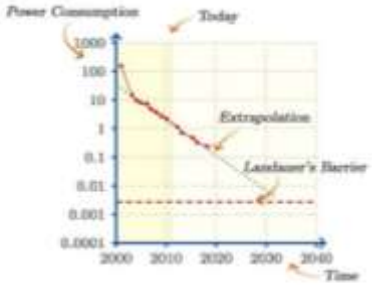
---

[411] Voir notamment Startup JITX Uses AI to Automate Complex Circuit Board Design de Evan Ackerman, juillet 2918, Using AI In Chip Manufacturing de Ed Sperling, août 2018, Machine learning for future systems designs par Elias Fallon, octobre 2020 et Using AI To Build Better Chips par Karl Freund, avril 2020. Voir aussi **Intendo Design** (2008, France) créé par Ramy Iskander, un ancien de la R&D de MentorGraphics et STMicroelectronics et qui propose une solution d'EDA (conception de circuits) qui exploite le machine learning présenté comme une EDA cognitive qui modélise la pensée du créateur humain.



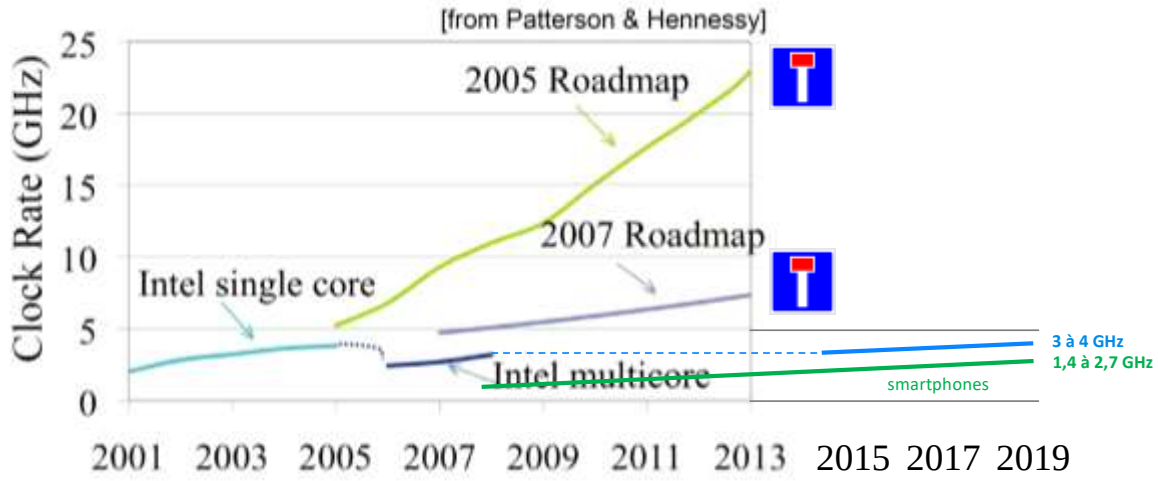


barrière de la chaleur
limite la fréquence d'horloge des processeurs



barrière de Landauer
limite basse de la consommation d'énergie d'un transistor

Les processeurs de smartphones lancés en 2007 tournaient de leur côté à une fréquence d'horloge d'environ 400 MHz (pour le premier iPhone) et atteignent aujourd'hui dans le meilleur des cas un peu moins de 3 GHz.

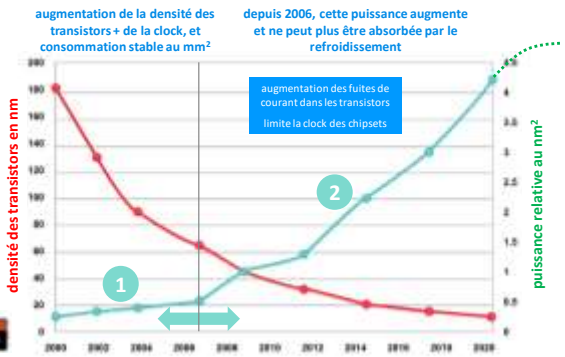


High Performance Computing - The Multicore Revolution, Andrea Marongiu, 2019 (41 slides)

Ceci est lié à la fin d'une autre règle, ou échelle, celle de **Dennard**. Edictée en 1974 par Robert Dennard, elle prévoyait que la consommation d'énergie des processeurs au mm^2 resterait constante au gré de la baisse de densité des transistors.

En gros, cela devait être permis par une baisse de la tension (en volts) et du courant (en ampère) linéaire par rapport à la dimension des transistors. Cette règle ou échelle (scale) a été vérifiée jusqu'en 2006. A ce moment-là, on a commencé à observer une augmentation de la consommation d'énergie des processeurs au mm^2 qui s'explique par des fuites de courant (leakages) dans les transistors.

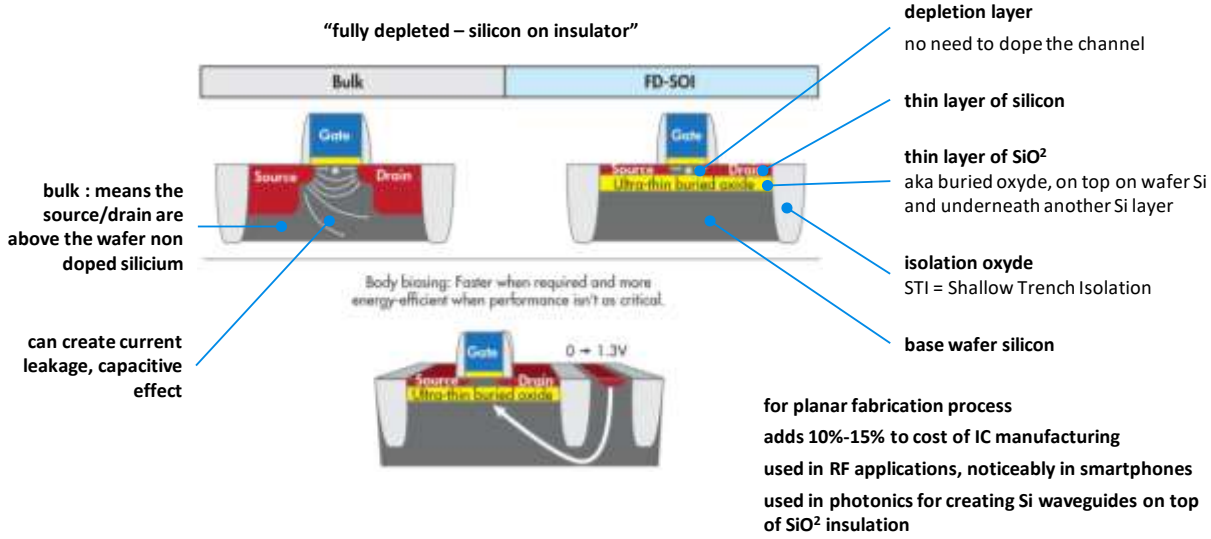
fin de la règle de Dennard en 2006



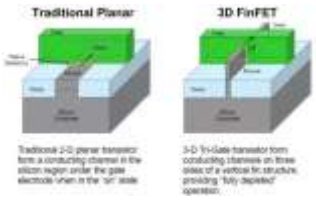
Cette course à la densité se poursuit cependant. Elle aboutit à un phénomène peu connu, celui du « dark silicon » lié au fait qu'à cause de cette densité énergétique galopante des processeurs, on ne peut pas les exploiter entièrement.

Des parades diverses ont été trouvées : réduire la fréquence d'horloge, utiliser des nœuds de processeurs tournant à des vitesses différentes selon les besoins, désactivation complète de nœuds, réserver la pleine charge à des laps de temps courts, créer des cœurs spécialisés optimisés par fonction, et enfin, améliorer les techniques de refroidissement des processeurs. Sachant dans le même temps que l'on cherche dans de nombreux usages de l'embarqué à utiliser un refroidissement passif. En pratique donc, il est rare qu'un processeur soit utilisé à pleine charge[412].

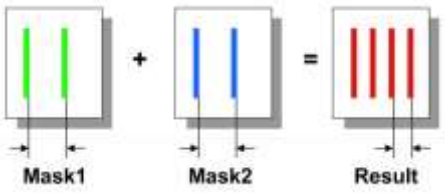
Les principales technologies d'intégration des 15 dernières années sont les transistors tri-gate « FinFET ». Avec eux, la densité est meilleure et il y a moins de fuites de courant et une baisse de consommation électrique. C'est encore mieux lorsque cette technologie est associée au FD-SOI issu du CEA-Leti, de SOITEC[413] et de STMicroelectronics, qui ajoute une couche d'isolant en oxyde de silicium (SiO_2) sur les wafers de silicium et isole bien les transistors les uns des autres (voir le schéma *ci-dessous*)[414]. Elle est notamment employée pour la fabrication de composants radio de smartphones tels que l'iPhone et en particulier pour supporter les bandes millimétriques de la 5G, au-delà de 20 GHz[415]. On le trouve aussi dans les chipsets neuromorphiques TrueNorth d'IBM. Le FD-SOI est adopté sous licence par Samsung en 28 nm ainsi que par Global Foundries en 22 nm.



Dans la course à la densité, le multi-patterning [416] permet d'affiner le dessin des transistors en alternant des tracés adjacents pour éviter qu'ils interfèrent entre eux lors de la gravure. Il coûte assez cher car il ajoute de nombreuses étapes à la fabrication des chipsets et peut augmenter le taux de rebus.



trigate transistors
less leaks, power saving
more complex steps



double and triple patterning
avoids small lines overlaps
but more process steps

[412] Voir au sujet de la fin de la loi de Moore : The End of Moore's Law, CPUs (as we know them), and the Rise of Domain Specific Architectures par John Hennessy, 2017 (28 slides) et The Decline of Computers as a General PurposeTechnology par Neil C. Thompson et Svenja Spanuth, novembre 2018 (66 pages).

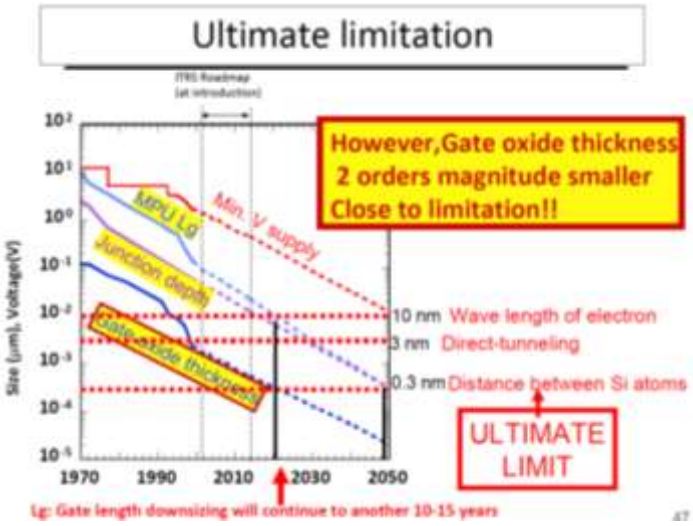
[413] SOITEC fait appel à de l'IA dans le contrôle qualité de ses wafers. Il utilise pour cela une technique d'analyse des images prises en infrarouge de ses wafers. Le traitement est réalisé dans le cloud sur AWS. Voir Soitec adopte l'IA en PaaS pour le contrôle qualité, mars 2019.

[414] Voir FD-SOI Technology par Bich-Yen Nguyen, Soitec, 2017 (63 slides).

[415] Voir Silicon Technologies for mmWave Applications par Ned Cahoon et al, de Global Foundries, 2019 (32 slides).

[416] J'explique le principe du multi-patterning dans A la découverte de la "fab" chez STMicroelectronics : 2, décembre 2014.

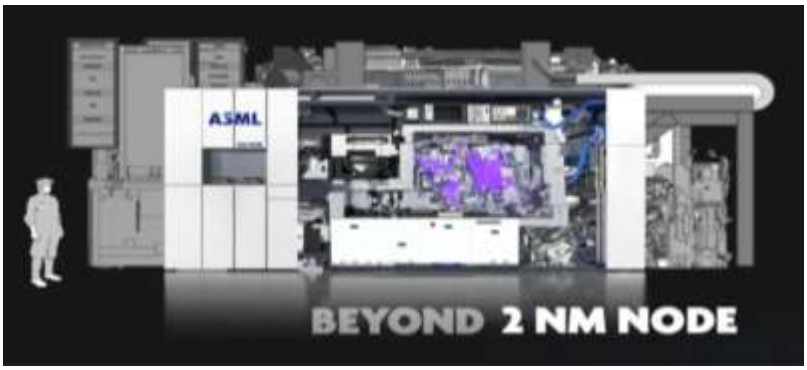
Au début des années 2000, des spécialistes considéraient qu'il n'y avait point de salut en-dessous des nœuds 20 nm[417].En 2009, Hiroshi Iwai du Tokyo Institute of Technology décrivait ces limites avec précision [418] (*ci-contre*), la limite ultime étant la taille d'un atome de silicium[419].



Or les chipsets en technologie 7 nm commençaient à être fabriqués dès 2018 chez Samsung [420] et TSMC pour les Qualcomm Snapdragon 855 et les Kirin 980 de la filiale HiSilicon de Huawei. Et les chipsets qui équipent les Huawei P40 et iPhone 12 de 2020 sont gravés en 5 nm chez TSMC.

Les roadmaps pour les prochaines années anticipent l'atteinte du 2 nm, en FinFET chez TSMC[421] ! A partir du 7 nm, la gravure en Extreme Ultra Violet (EUV) devient indispensable du fait de la meilleure résolution spatiale qu'elle procure. Elle permet en plus d'éviter le multi-patterning et de réduire le nombre d'étapes de fabrication des chipsets de plus d'une centaine de masques à environ 80 masques[422]. Les machines de gravure EUV d'**ASML** sont des concentrés de technologie associant une mécanique de précision capable de déplacer rapidement les wafers sous l'optique d'exposition avec une précision inférieure au nanomètre et des optiques spécifiques dotées d'abberrations réduites au minium. Les rayons extrême ultra-violet sont générés par des lasers illuminant un plasma sous vide de micro-goutellettes d'étain fondu ! Ces machines coûtent jusqu'à $250M l'unité et il en faut environ 4 à 5 pour une unité de production classique [423]!

Ils en livrent actuellement une trentaine par an. ASML est l'un des rares leaders européens des technologies numérique. Il est en situation de quasi-monopole dans les équipements de lithogravure de semiconducteurs, et le seul à fournir des machines de gravure en EUV. Leur chiffre d'affaire 2020 était de $14B, la moitié réalisée chez TSMC !



---

[417] En voici un exemple avec Why is CMOS scaling coming to an END ? de Nor Zaidi Haron et Said Hamdioui, 2006 (6 pages).

[418] Voir Downsizing of transistors towards its Limit, Hiroshi Iwai du Tokyo Institute of Technology, 2009 (79 slides).

[419] L'excellent dossier After Moore's Law, paru dans The Economist en mars 2016, détaillait bien la question en expliquant pourquoi la loi de Moore des transistors CMOS pourrait s'arrêter en une douzaine d'année lorsque l'on descendra au niveau des 5 nm d'intégration.

[420] Voir Samsung Foundry Roadmap: EUV-Based 7LPP for 2018, 3 nm Incoming de Anton Shilov, mai 2018.

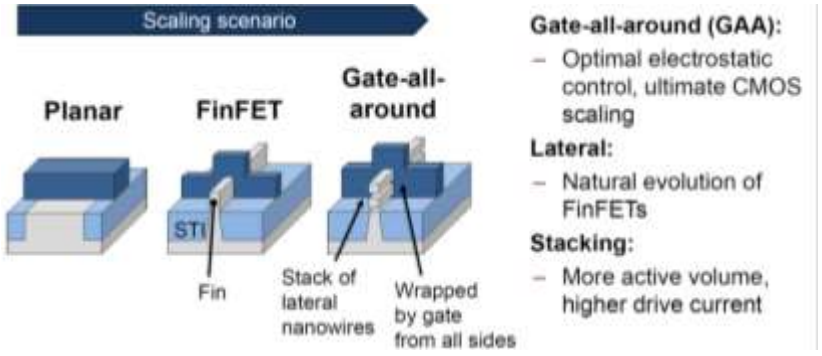
[421] Voir Where are my GAA-FETs? TSMC to Stay with FinFET for 3nm par Dr. Ian Cutress, août 2020.

[422] Voir The Impact of EUV on the Semiconductor Supply Chain, 2018 (19 slides).

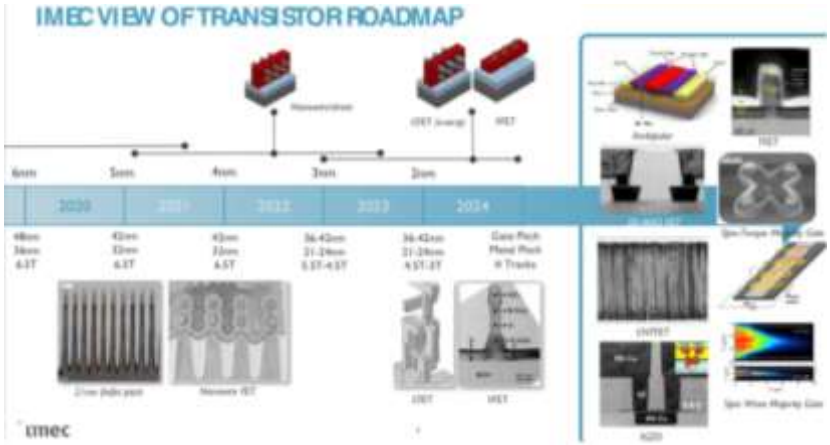
[423] Voir la description de la technologie et de la roadmap d'ASML dans Enabling Semiconductor Innovation and Growth - EUV lithography drives Moore's law well into the next decade, 2018 (37 slides).

Cette montée en flèche des coûts explique les montants exhorbitants à investir pour des fabs de semiconducteurs descendant à un niveau d'intégration inférieur à 10 nm, qui sont de plus de $15B. Sachant qu'il faut en plus des dizaines d'autres machines dont le coût se chiffre aussi en millions et dizaines de millions de dollars l'unité.

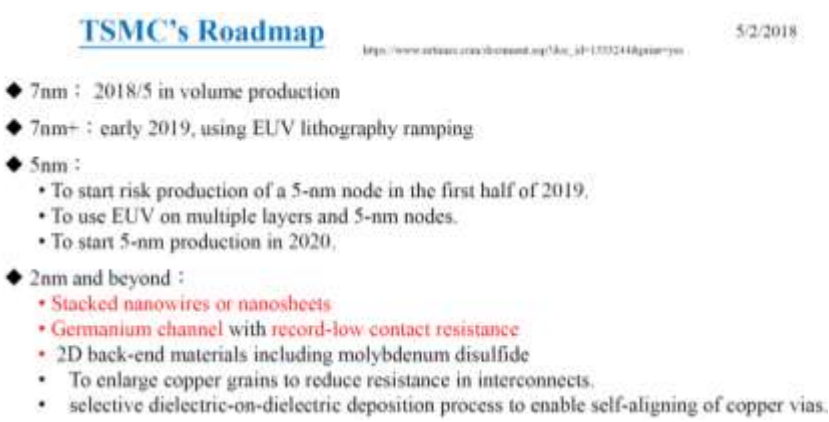
L'une des techniques étudiées pour faire croître la densité des transistors est celle des « Gate-all-around » (GAA) et des nano-fils qui permettent d'améliorer l'intégration verticale des transistors[424]. Sont aussi envisagés des « nanosheets » qui remplacent les fils par des pico-feuilles ultra-fines.



On retrouve ce genre de technologie avec d'autres (TFET, CNTFET, etc) dans la roadmap de Chee Wee Liu[425] d'où est issue le schéma *ci-contre*.



Elles permettraient de descendre à une densité de 2 nm d'ici 2024 grâce à l'emploi de ces transistors encore plus denses et intégrés verticalement. La roadmap du fondeur Taïwainais **TSMC** prévoit même l'au-delà du 2 nm[426]!



TSMC fabrique les chipsets de Huawei (modulo les embargos américains), d'Apple, Qualcomm et Nvidia jusqu'en technologie 5 ou 7 nm FinFET, selon les cas.

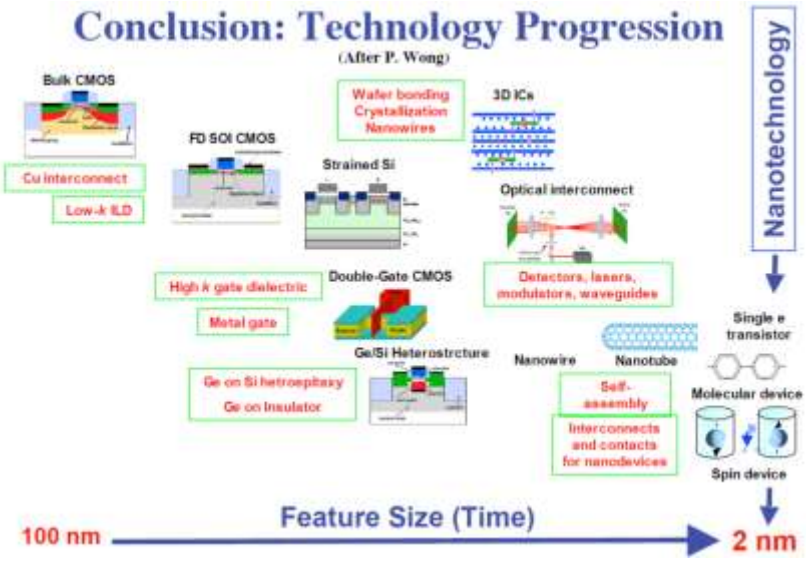
C'est l'un des deux derniers géants mondiaux qui subsistent dans la course à l'intégration avec Samsung. Leur CA 2020 était de $48B.

[424] Le schéma est issu de la présentation From Gate-all-around MOSFETS based on vertical stacked horizontal Nanowires de Hans Mertens, 2017 (78 slides).

[425] Voir Innovations Enabling Semiconductor Roadmap du Taïwanais Chee Wee Liu, 2018 (45 slides). Pour mémoire, 1 nm = 10 atomes de silicium.

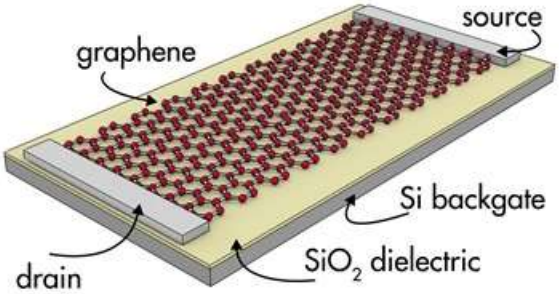
[426] Voir Performance and Design Considerations for Gate-All-around Stacked-NanoWires FETs, 2017 (47 slides).

D'autres techniques sont envisagées à plus long terme comme des nanotubes de carbone comme chez **Nantero**[427] avec sa NRAM au standard DDR4 et proposée sous licence[428].



En-dessous de 2 nm, il faudra peut-être commencer à faire une croix sur la loi de Moore. Les architectures multi-cœurs atteignent de leur côté leurs limites car les systèmes d'exploitation et les applications sont difficiles à ventiler automatiquement sur un nombre élevé de cœurs, au-delà de 4.

Du côté de la vitesse d'horloge, des pistes sont explorées avec du graphène. IBM avait annoncé en 2011 avoir produit des transistors au graphène capables d'atteindre une fréquence de 155 GHz, et en gravure à 40 nm[429]. Une performance en laboratoire d'aboutit pas toujours à de l'industrialisation ! Cela peut-être lié à une difficulté à fabriquer le composant avec un taux de défaut raisonnable.



Il faudra encore patienter un peu de ce côté-là même si cela semble très prometteur et avec des débouchés dans tous les domaines et pas seulement dans l'IA[430].

Le CEA-Leti à Grenoble développe pour sa part la technologie CoolCube d'empilement en 3D de chipsets les uns sur les autres utilisant un procédé d'assemblage à relativement basse température. Un dérivé de cette technologie utilisant un interposeur en silicium de séparation entre chipsets empilés devrait être utilisé par Intel suite à un partenariat annoncé avec le CEA-Leti en octobre 2020[431].

Alors, la loi de Moore est foutue ? Pas si vite ! Elle avance par hoquets et il reste encore beaucoup de mou sous la pédale pour faire avancer la puissance du matériel. Mais tout ce que nous venons de voir concerne principalement que la densité des processeurs. C'est loin d'être suffisant pour accélérer les logiciels de l'IA !

La principale voie choisie pour améliorer la performance de l'IA consiste à créer des processeurs qui exécutent nativement les primitives mathématiques des réseaux de neurones du deep learning. C'est le meilleur moyen permettant de paralléliser les traitements au maximum, de les accélérer, et de réduire la consommation d'énergie associée.

---

[427] Voir Carbon Nanotube DRAM, de Susan Rambo, août 2018.

[428] Le schéma est issu de How far can we push Si CMOS What lies beyond de Krishna Saraswat, Stanford University (27 slides).

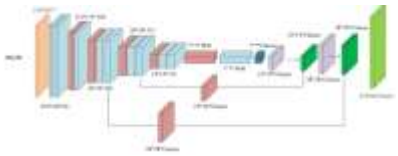
[429] Voir High-frequency, scaled graphene transistors on diamond-like carbon par Yanqing Wu et al, Nature, 2011 (6 pages).

[430] La recherche mondiale produit nombre de papiers concernant la creation de transistors au graphène. Par exemple : Large spin-relaxation anisotropy in bilayer-graphene/WS2 heterostructures, par S. Omar & Al, octobre 2019 (13 pages).

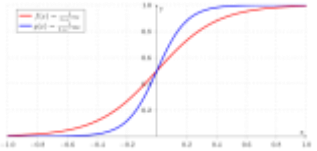
[431] Voir CEA-Leti Announces Collaboration with Intel to Advance Chip Design Through Cutting-Edge 3D Packaging Technologies, octobre 2020. Voir aussi Le CEA-Leti développe un processeur avec chiplets et interposeurs actifs, mai 2020.

Plusieurs types de calculs doivent être réalisés par ces processeurs :

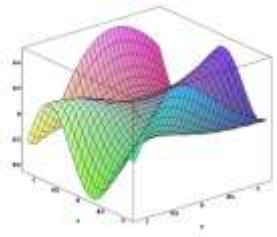
- Des **multiplications de matrices et de vecteurs**, en particulier dans les réseaux de neurones convolutifs de reconnaissance d'images[432]. La gestion de « tenseurs » correspond à cette fonction. Elle permet d'exécuter des calculs plus rapidement qu'avec des ALU de GPU (Arithmetic Logic Units) qui parallélisent bien les calculs sur des dizaines, centaines et milliers d'ALU mais moins efficacement que des multiplicateurs et additionneurs de matrices du fait d'allers et retours plus fréquents avec la mémoire (cache ou RAM). Ces calculs peuvent être éventuellement optimisés pour des vecteurs et matrices qui sont parfois remplis d'une grande quantité de zéros (les « sparsed matrixes »).
- La capacité à gérer des **fonctions non linéaires** comme les sigmoïdes, softmax, tanh et autres fonctions de normalisation dans les neurones, et si possible avec une grande précision en nombres flottants, surtout lors de l'entraînement des réseaux de neurones[433].
- Le calcul de **fonctions dérivées** pour la gestion de la descente des gradients lors de l'entraînement de réseaux de neurones. Là encore, une grande précision est préférable pour les phases d'entraînement. Sans compter l'épineuse question de la recherche aussi rapide que possible de minimums globaux lors de ces descentes de gradients.
- Ces calculs sont plutôt réalisés en **nombres flottants** sur serveurs avec une grande précision pour l'entraînement et exécutés ensuite en **nombres entiers** dans les systèmes embarqués, comme les smartphones, pour économiser de l'énergie et accélérer les calculs.
- Le tout doit être réalisé en faisant en sorte que l'**accès à la mémoire** qui contient les paramètres des réseaux de neurones soit le plus rapide possible. Les architectures rivalisent donc d'ingénuité pour rapprocher la mémoire des unités de calcul dans les chipsets du marché.



multiplications et additions
matrices et de vecteurs



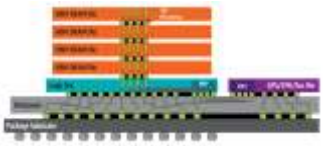
fonctions non linéaires
sigmoïdes, tanh, softmax



fonctions dérivées
descente de gradient

FP16, FP32, INT4, INT8, INT16

calculs
en flottants (entraînement)
ou en entiers (inférence)



accès rapide
à la mémoire

limiter la consommation d'énergie

Une alternative consiste à gérer des **neurones matériels** avec leurs entrées et sorties, fonctions de calcul et mémoire internes, ces neurones étant organisés en réseaux maillés interconnectés. C'est la forme la plus « pure » de chipset neuromorphique.

---

[432] Voir à ce sujet A Survey of Accelerator Architectures for Deep Neural Networks par Yiran Chen, janvier 2020 (11 pages).

[433] Ces fonctions utilisent des exponentielles. Comment ces exponentielles et les fonctions trigonométriques sont-elles réalisées dans les processeurs ? Sur les chipsets serveurs, elles sont calculées par des approximations polynomiales. Dans l'embarqué, elles peuvent l'être avec de simples tables de correspondance et des approximations linéaires. Voir par exemple Efficient Double-Precision Cosine Generation par Derek Nowrouzezahrai et al, 2005 (5 pages).

Elle est notamment favorisée pour les systèmes fonctionnant par apprentissage par renforcement dans des systèmes embarqués. On la retrouve mise en œuvre dans les processeurs exploitant des neurones à impulsions (spiking neurons).

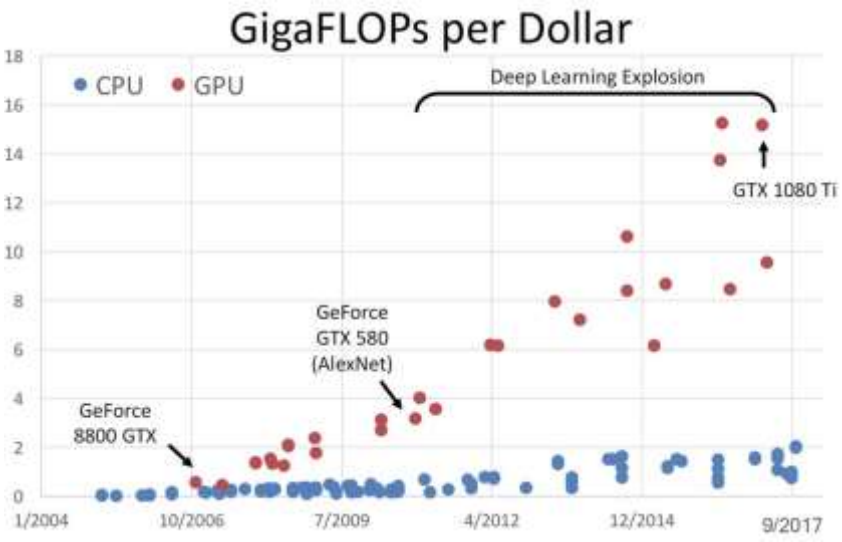
La loi de Moore est censée s'appliquer à des solutions commercialement disponibles, et si possible, en volume. Or ce n'est pas toujours le cas. Ainsi, l'évolution de la puissance des supercalculateurs est mise en avant comme un progrès technique validant la loi de Moore. Or, ces calculateurs sont créés avec des moyens financiers quasiment illimités et n'existent qu'en un seul exemplaire, souvent réalisé pour de la recherche militaro-industrielle ou de grands projets de recherche fondamentale (aérospatial, génomique, météo).

Dans la plupart des cas, ces technologies "de luxe" font leur apparition dans des produits grand public quelques années ou décennies plus tard. Ainsi, la puissance des super-calculateurs des années 1990 est-elle apparue dans les consoles de jeu des années 2000. Au lieu de faire des calculs en éléments finis pour des prévisions météo, les consoles de jeux calculent des millions de polygones pour générer des images de synthèse en 3D temps réel.

Mais cette puissance n'est pas homothétique dans toutes les dimensions. Si la puissance de calcul est similaire, les capacités de stockage ne sont pas forcément les mêmes. Les volumes de données gérés par les supercalculateurs de cette époque étaient déjà de plusieurs ordres de grandeur supérieurs à ceux des smartphones d'aujourd'hui.

GPGPU

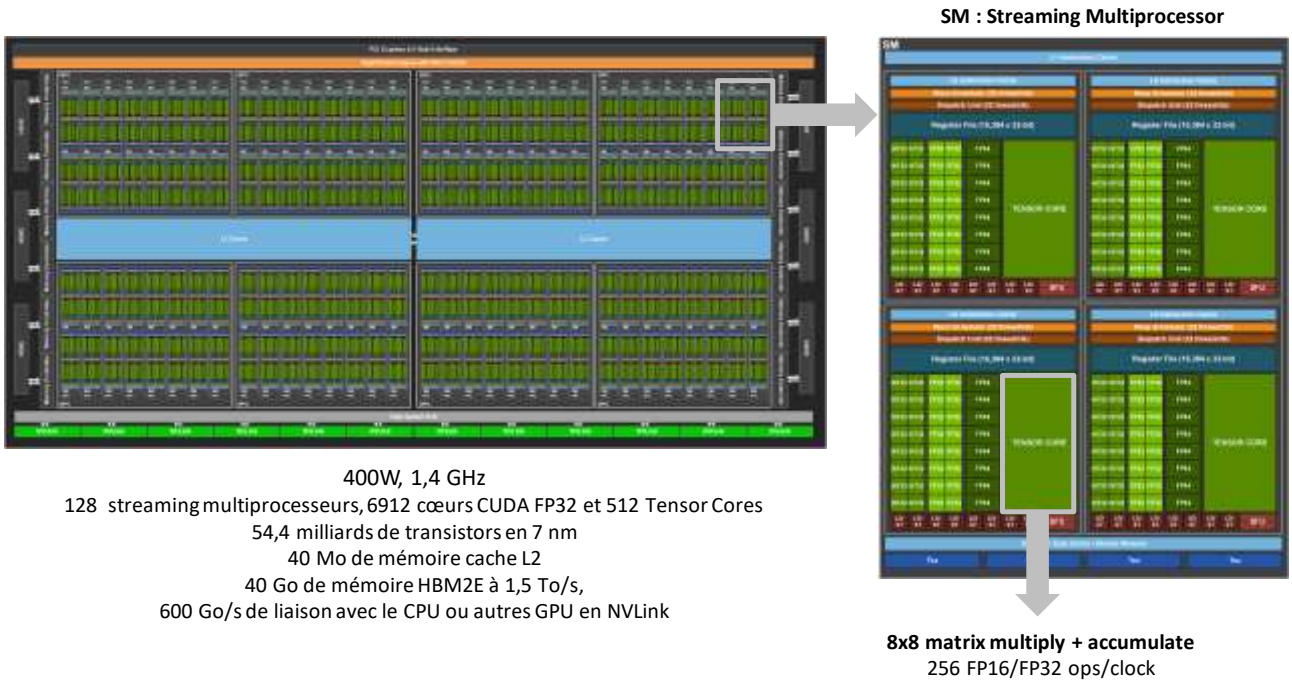
Les GPGPU [434] constituent la solution matérielle la plus largement déployée pour accélérer les réseaux de neurones et le deep learning. Ce sont eux qui ont rendu possible le deep learning, notamment pour le traitement de l'image et à partir de 2012 avec AlexaNet. On le voit dans le schéma *ci-contre* qui décrit la baisse drastique du coût au GigaFlops par dollar. La puissance brute a également augmenté d'autant sur le même laps de temps.



Le leader de ce marché est l'américain **Nvidia** qui fournit une vaste gamme de chipsets, de cartes et de serveurs équipés de GPGPU adaptés aux traitements de l'IA. Ils sont couramment installés dans des data centers et dans le cloud. Leur challenger AMD est à la peine, avec ses GPU et son API OpenCL qui bénéficie d'un support plus que médiocre par l'écosystème logiciel du deep learning[435].

[434] GPGPU : General Purpose GPU. A opposer aux GPU qui sont dédiés aux fonctions graphiques, notamment pour l'exécution de jeux vidéos.

[435] Voir AMD Too Late To NVIDIA Deep Learning Party, de Boris Suvorov, décembre 2016. Lancée fin 2018, la carte AMD Radeon Instinct MI50 Accelerator est dotée de 16 Go de mémoire HBM2 avec 1 To/s de bande passante. Elle est gravée en 7 nm, tourne à 1,725 GHz avec 60 unités de traitement et 3840 stream processors. Elle consomme 300W et génère 53 TOPS en entiers 8 bits. Son bus est en PCI 4.0 et la communication entre ces GPU est de 184 Go/s via AMD Infinity Fabric, qui est leur équivalent de NVLink de Nvidia.



Le GPGPU phare de Nvidia depuis son lancement en 2017 était le V100, destiné aux serveurs et aux supercalculateurs. Ses 21,1 milliards de transistors gravés en 12 nm totalisent de 120 Tflops[436] ! Jusqu'à présent, les GPU comprenaient une myriade de cœurs, les ALUs (Arithmetic & Logic Units), à même d'effectuer des opérations mathématiques simples (multiplications, divisions, additions, soustractions). Les logiciels utilisant l'interface CUDA répartissaient les traitements dans le GPU pour les paralléliser comme pour la génération des effets graphiques 2D et 3D. Pour le deep learning, les calculs étaient aussi répartis dans ces cœurs mais ce n'était pas optimal. Les V100 comprennent des « tensor cores » qui sont des multiplicateurs et additionneurs de matrices de 4x4 en nombres flottants 16 bits permettant de mieux paralléliser les traitements d'un réseau de neurones, surtout dans les réseaux convolutifs[437].

Les V100 comprennent 80 « streaming multiprocessors », comprenant un total de 5120 cœurs CUDA traditionnels (avec 64 cœurs en flottant 32 bits, 32 cœurs flottant 64 bits et 64 cœurs entier par SM) et 640 « tensor cores » (8 par SM). Cette architecture flexible et générique s'adapte à de nombreux types de traitements. Elle est par ailleurs très bien supportée côté logiciels et frameworks. L'architecture des V100 est hybride avec ses ALU exploitables pour toutes sortes de calculs et les tenseurs faits pour le deep learning qui occupent environ 30% de la surface du processeur.

Les composants V100 comprennent la puce du GPGPU et des puces de mémoire au standard HBM2 de 16 ou 32 Go avec un débit extraordinaire de 1800 Go/s[438]. Cette quantité de mémoire est critique car elle conditionne le nombre de paramètres qu'un réseau de neurones pourra exploiter. On atteint très vite cette quantité de mémoire avec un réseau convolutif classique. Cela explique, entre autres choses, pourquoi la résolution des images traitées dans ces réseaux de neurones est souvent limitée à 227 ou 224 pixels de côté.

---

[436] Voir pas mal de détails dans NVIDIA Volta Unveiled: GV100 GPU and Tesla V100 Accelerator Announced. A noter que Nvidia entretient une équipe de développeurs en France sous la responsabilité de Julien Demouth qui participe à la conception de ses GPU pour le deep learning. Une nouvelle génération gravée en 7 nm pourrait être annoncée par Nvidia en 2020. Voir aussi Volta Tensor Core Training, ORNL, août 2019 (113 slides) et Introduction to mixed precision training par Dusan Stosic, 2019 (78 slides).

[437] Nvidia développe aussi des chipsets de recherche de la série RC, le dernier étant le RC18. Voir Nvidia Shows off Tech Chps with R18 Inference Chip par Timothy Prickett Morgan, septembre 2019. C'est un chipset d'inférences présentant la particularité de générer un record de 9.5 tops par watt consommé, pour un total de 128 tops. Il manipule des matrices de 8x8.

[438] Voir High-Performance Memory Challenges de Ann Steffora Mutschler, avril 2018, qui explique les défis de la gestion de la mémoire dans les GPU dédiés aux applications de l'IA.

Un GPGPU n'est utilisable pour des réseaux de neurones que si la mémoire est facilement partagée entre les cœurs de GPU[439]. C'est ce que propose justement Nvidia avec son architecture GPUDirect RDMA et avec son bus NVLink qui atteint la vitesse de 600 Go/s depuis 2020.

En mai 2020, Nvidia annonçait le successeur des V100 avec la série Ampere A100. Cette nouvelle famille de GPGPU est gravée par TSMC en 7 nm sur une surface de 826 mm^2, contre une gravure de 12 nm et 815 mm^2 pour les V100. Le processeur A100 comprend 54 milliards de transistors avec des cœurs tenseurs de troisième génération. Le nombre de cœurs CUDA passe de 5120 à 6912. L'augmentation de performance mise en avant va de x2,5 à x20 selon les cas d'usage et la précision des calculs. Cette amélioration provient notamment de l'augmentation de la capacité des tenseurs, au nombre de 512, qui peuvent gérer parallèlement 256 opérations pour faire des multiplications et additions de matrices 8x8. La mémoire HBM2 passe de 32 à 40 Go, avec une version 80 Go annoncée mi-novembre 2020, permettant d'entraîner des modèles de deep learning encore plus grands[440]. La mémoire cache intégrée augmente aussi pour passer à 40 Mo[441].

Comme le V100, le A100 vise aussi les supercalculateurs et le calcul scientifique, avec une capacité de calcul de 19,5 TFLOPS en nombres flottants 64 bits.

La hiérarchie des unités de traitement du GA100 comprend 8 GPC (GPU processing clusters) qui contiennent chacun 8 TPC (texture processing clusters) qui contiennent chacun 2 SM (streaming multiprocessor). Ce qui fait un total de 128 streaming processors. Mais, nuance, le processeur est censé être commercialisé en deux versions : le GA100 avec 128 SM et 512 tensor cores et le A100 avec 108 SM et 432 tensor cores.

Autre différence dans les tensor cores : dans le V100 de 2017, le calcul matriciel côté multiplication se faisait sur un flottant 16 bits, et l'addition avec une troisième matrice du flottant 32 bits. Il se trouve que le "mixed precision deep learning training" est apprécié car il est très efficace en ressources machine utilisées.

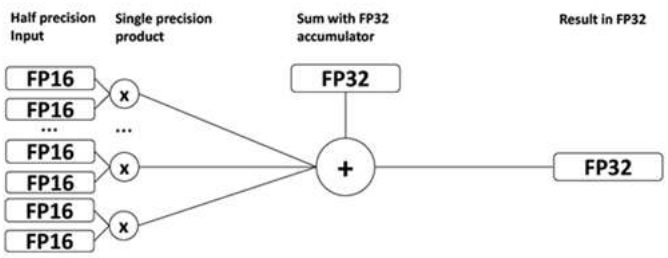


Fig. 3: FMAs in NVIDIA Tensor Cores.

Dans le A100, la multiplication matricielle est aussi réalisable en nombre flottant 32 bits, ce qui peut améliorer la précision… si besoin est. Cela devrait être plus utile pour le calcul scientifique haute performance qui a généralement besoin de précision plus que pour l'entraînement de réseaux de neurones où l'on cherche à s'en passer.

Ces tenseurs font du FMA pour "fused multiply-add" : une multiplication entre deux matrices puis une addition avec une troisième. On passe d'une architecture qui associait deux modules gérant 64 opérations simultanément (4x4x4 fois deux, en haut à gauche du schéma ci-dessous) à un nouveau qui gère 256 opérations d'un coup, avec une flexibilité d'organisation qui est pilotée par logiciel via la bibliothèque CUDA 11.

[439] Voir How To Build and Use a Multi GPU System for Deep Learning de Tim Detmers, 2014.

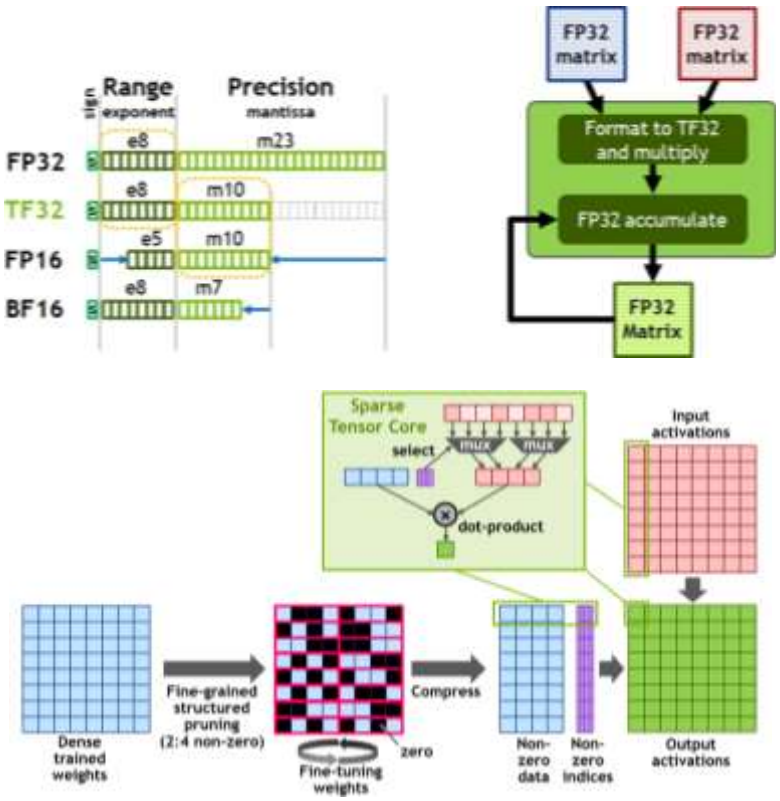
[440] Voir NVIDIA Announces A100 80GB: Ampere Gets HBM2E Memory Upgrade par Ryan Smith, novembre 2020.

[441] Voir NVIDIA Ampere Architecture In-Depth | NVIDIA Developer Blog par Ronny Krashinsky, mai 2020.

Le A100 supporte un nouveau type de donnée, le TF32 pour Tensor Float 32. Il est exploité dans les multiplications de matrices avec la mantisse de 10 bits du FP16 (nombres après la virgule) et l'exposant 8 bits du FP32 (cf ci-dessus). Cela permet de faire des calculs avec une bonne précision et moins de charge que du calcul flottant 32 bits complet (FP32).

Nvidia a aussi intégré un système d'optimisation du calcul matriciel qui permet de gérer des "sparse matrix", des matrices partiellement remplies dites clairsemées ou matrices creuses qui sont courantes dans les réseaux de neurones et en particulier dans ceux qui traitent du langage.

Tout cela génère des gains de performance variés expliqués ci-contre.



Pour terminer sur les capacités du GPGPU A100, passons au côté mémoire et gestion des entrées/sorties :

- La mémoire vive intégrée évolue de 16 Go (2017) / 32 Go (2018) à 40 Go au standard HBM2E entre le GV100 et le GA100. C'est une augmentation modeste.
- Le débit de l'accès à la mémoire passe de 900 Go/s à 1555 Go/s grâce au standard HBM2E. Ces puces de mémoire HBM2E sont notamment proposées par les coréens SK Hynix et Samsung.
- On passe par contre de 6 à 40 Mo pour la mémoire cache L2 (intermédiaire, entre la mémoire HBM2E et le cache L1) ce qui est intéressant pour remplir plus rapidement les matrices de calcul.
- Le bus pour la carte comprenant le processeur passe au PCI Express 4.0 qui double la bande passante à 31,5 Go/s. pour une liaison x16 (que l'on peut agréger). Donc, cela double la vitesse d'accès aux supports de stockage SSD, ce qui permettra d'accélérer aussi de ce côté-là l'entraînement d'un réseau de neurones devant scanner une grande base d'entraînement.
- La liaison NVLink inter GPU/CPU supporte un débit 600 Go/s vs 300 Go/s avec le GV100. Elle est utilisée au sein de serveurs comme le HGX de Nvidia.
- Le GPU supporte aussi des interfaces réseaux allant jusqu'à 200 Gbits/s au standard InfiniBand, qui est supporté par les cartes Mellanox de Nvidia.

Tout cela permettait à Nvidia de publier des benchmarks impressionnants battant quelques records[442].

Une dernière fonction clé est celle de la **virtualisation** (MIG = Multi-Instance GPU) qui s'appuie sur un partitionnement du processeur sur 7 instances et est très utile pour partager les ressources du GPU pour des services en cloud.

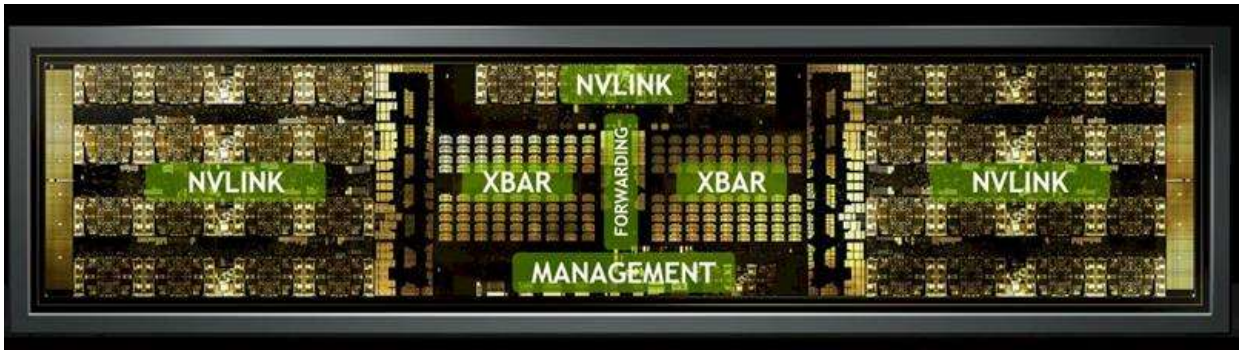
[442] Voir New AI Inferencing Records par Samuel K. Moore, octobre 2020.

Les GPGPU Nvidia sont intégrés dans une gamme de produits divers allant de la station de travail aux supercalculateurs.

- La **DGX Station** est une station de travail dotée de quatre V100 et comprenant aussi un CPU Intel Xeon plus 128 Go de RAM HBM2 pour les GPU et 256 Go de RAM pour le CPU. Elle tourne sous Ubuntu. Une version équipée du A100 Ampere était annoncée fin 2020.
- Des serveurs Nvidia **DGX A100** de 5 PFLOPS (évalués en FP16) intégrant 8 A100 complétés de 15 To de SSD, de neuf cartes réseaux Mellanox ConnectX-6 VPI de 200 GBits/s et de deux CPU AMD EPYC 7742 Rome, et qui remplacent des processeurs Intel des serveurs DGX-1 et DGX-2. Notamment parce qu'ils sont les premiers à supporter le bus PCI 4.0 qui est plus rapide que le 3.0, en particulier pour les accès SSD.

Nvidia avait acquis **Mellanox** en 2019. Nous avons donc ici une intégration de son offre de connectivité réseau très haut débit dans l'ensemble de l'architecture de serveur Nvidia qui prend tout son sens. Celle-ci est point à point avec de nombreux serveurs, permettant une répartition optimale des traitements distribués, un point clé pour le calcul haute performance.

Ces serveurs succèdent au DGX-1 lancé en 2017 et faisant un PFLOPS avec8 V100 et au DGX-2 faisant 2 PFLOPS avec16 V100. Ces derniers exploitaient une nouvelle interconnexion NVLink 2.0 avec des chipsets dédiés, les NVSwitch (*ci-dessous*), avec 2 milliards de transistors, 18 ports NVLink 2.0 supportant 25 Go/s de communication bi-directionnelle par port donnant un total de 900 Go/s[443]. Ils sont ausi équipés de connectique Infiniband Mellanox (maintenant intégrée à Nvidia)[444].



[443] Voir Inside Nvidia's NVSwitch GPU Interconnect et Building biger, faster GPU clusters using NVSwitches tous deux publiés en avril 2018 et Nvidia DGX-2 Details at Hot Chips 30 par Patrick Kennedy, août 2018. A noter : la génération de chipsets serveurs Power9 d'IBM supporte déjà NVLink.

[444] L'interconnectivité entre CPU et GPU est un facteur clé de performance.Voir Performance Analysis of Deep Learning Workloads on Leading-edge Systems par Yihui Ren et al, octobre 2019 (11 pages) qui analyse la performance à l'entraînement de différentes architectures on-premise et en cloud à base de GPU Nvidia.

- Nvidia a un écosystème de partenaires qui fournissent des serveurs équipés de A100 ou V100. Des serveurs de constructeurs comme **HPE** (ProLiant DL380 Gen10 Server et Apollo 6500 Gen10 System), **Dell**, **Fujitsu**, **Quanta**, **Lenovo**, **Gigabyte** et **SuperMicro** intègrent le A100. Des offres de serveurs intégrant le V100 avaient auparavant été lancées par **PNY** (1985, USA), **Lambdalabs**, **Cisco** (complétés de routeurs maison[445]), **IBM** (avec ses AC922 Power Systems qui comprennent deux CPU Power9 et des V100[446]), **Atos** (qui lançait en 2017 ses serveurs hybrides BullSequana S qui associent des CPU Intel, des GPU Nvidia V100 et des extensions à base de FPGA, le tout permettant d'optimiser ses calculs dans l'IA[447]).

Nvidia DGX A100
8 GPU A100
5 petaflops, $400K

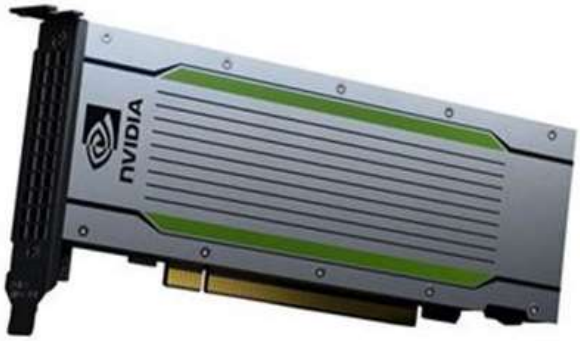
Cisco UCS 240 ML M5
4 GPU A100 + Intel Xeon

PNY A3I
2 DGX A100
10 petaflops
DDN A^{3}I storage

Atos BullSequana S
2 à 32 CPU Intel Xeon Phi et Nvidia V100

- Pour le stockage, **NetApp** propose une architecture de référence, ONTAP, qui comprend quatre DGX-1, un système de stockage Netapp de 95 To et un routeur Cisco 100 Gbits/s[448]. **Dell-EMC** fait de même avec une combinaison d'un système de stockage Isilon F800 all-flash Dell EMC combiné à un serveur Nvidia DGX-1.

- En septembre 2018, **Nvidia** lançait une carte GPU au format PCIe de 75W pour serveurs dédiée à l'exécution de réseaux de neurones entraînés et intégrant un GPU Tesla T4 exploitant l'architecture Turing apparue dans les GPU lancés au printemps 2018. Ces cartes Tensor Hyperscale Inference Platform intègrent maintenant des serveurs de datacenters. Le GPU Nvidia Tesla T4 comprend 320 noyaux Turing Tensor Cores (multiplicateurs de matrices 4x4) et 2560 unités de traitement ALUs CUDA classiques.



  Les calculs sont réalisés avec des précisions FP16 et FP32 (16 et 32 bits en nombre flottant) ainsi que INT8 et INT16 (entier sur 8 et 16 bits). Il atteint 65 TFLOPS en FP16, 130 téraops en INT8 et 260 TOPS en INT4. Le GPU est supporté par la bibliothèque logicielle TensorRT 5 qui comprend un optimiseur d'inférence et un runtime supportant les frameworks TensorFlow, MXNet, Caffe2, Matlab et les frameworks supportés par ONNX comme ceux de Microsoft.

---

[445] Voir Cisco UCS 480 ML M5 Server – Performance and Capacity for AI, septembre 2018.

[446] Voir IBM Bookends AI With New Accelerated Power9 System par Timothy Prickett Morgan, janvier 2020.

[447] Voir Atos se projette dans l'intelligence artificielle embarquée par Florian Dèbes, mai 2019.

[448] Voir NetApp ONTAP AI, Powered by NVIDIA, mars 2019 (31 pages).

- Ajoutons **Run.AI** (2018, Israël, $43M), une startup qui propose une solution de gestion de virtualisation des ressources de calcul de GPGPU Nvidia dans le cloud dans des containers Kubernetes.

Côté cartes, Nvidia lançait aussi en 2020 un GPU pour carte graphique, le GA102 intégré dans les GeForce RTX 3090 et RTX 3080, le premier étant la plus puissante carte d'accélération GPU en date de Nvidia. Elle supporte les jeux vidéo 8K en HDR et 60 images par seconde. Le GA102-300 de la carte GeForce RTX 3090 comprend 10496 cœurs CUDA, 24 Go de mémoire GDDR6X, et le système d'upscaling utilisant les tenseurs du GPU, le nouveau DLSS 8K. Il est gravé en 8 nm chez Samsung et intègre 28 milliards de transistors, 328 tenseurs, soit environ la moitié du GA100. La carte consomme 350W et était lancée à $1500. Le GA102 se distingue des GA100 par l'ajout des « RT Core », des modules qui servent à réaliser du ray tracing pour les jeux vidéo et ajoutent une fonctionnalité de création de flous de mouvements[449].

Nvidia est aussi présent dans les systèmes embarqués avec :

- La **Nvidia Drive AGX Pegasus** qui supporte la conduite entièrement autonome de niveau 5, Elle a une puissance de 320 TOPS. Elle exploite quatre processeurs embarqués dont deux de la série Xavier totalisant 9 milliards de transistors (*ci-dessous*). C'est un chipset différent du V100 car il est plus généraliste, comme un chipset de smartphone. Il ne comprend que 20 tenseurs contre 640 dans le V100 ce qui est normal car ce genre de chipset sert à l'inférence de réseaux de neurones et pas à leur entraînement. La carte s'interface avec 16 capteurs haut-débit donc des capteurs ultra-sons, caméras, radars et LiDARs. Elle est dotée de plusieurs connecteurs Ethernet 10 Gbits/s. Sa bande passante mémoire excède 1 To/s. Cette carte permet l'exécution de modèles de deep learning entraînés sur des serveurs Nvidia DGX-1 ou DGX-2[450].

  En octobre 2020, Nvidia annonçait le successeur des Xavier avec Orin, qui est complété de cœurs Ampere[451]. Le haut de gamme est une carte Drive Robotaxi dédiée à la conduite autonome de niveau 5 intégrant deux SoC Orin et deux GPU Ampere GA100, pour 800W et 2000 TOPS.

- Les cartes **Jetson** font partie d'une gamme de cartes destinées à des applications diverses : systèmes de vidéo surveillance, véhicules autonomes ou à conduite assistée et objets connectés en tout genre. Un système de surveillance peut ainsi disposer de son intelligence locale pour n'envoyer vers les serveurs que des alertes consommant peu de bande passante et via des réseaux télécoms de type LPWAN (Low Power Wide Area Network) comme celui de Sigfox. En février 2018, **Nvidia** annonçait un partenariat avec **AnyVision** pour créer des caméras de surveillance intégrant de la reconnaissance faciale. Cela s'intègre dans l'initiative Metropolis de Nvidia qui est destinée aux Smart City qui associe également **Cisco**, l'intégrateur de systèmes de vidéosurveillance **Genetec**, le spécialiste du machine learning **Omni AI** et **MotionLoft** qui propose une solution logicielle de comptage de personnes et de véhicules dans l'espace public.
- La carte **Jetson AGX Xavier** est destinée aux véhicules. Elle comprend un GPU Volta avec 512 cœurs et 64 tenseurs, délivrant 11 TLOPS (en flottants 16 bits) ou 22 TOPS (en entiers 8 bits) et un CPU Arm 8 cœurs.

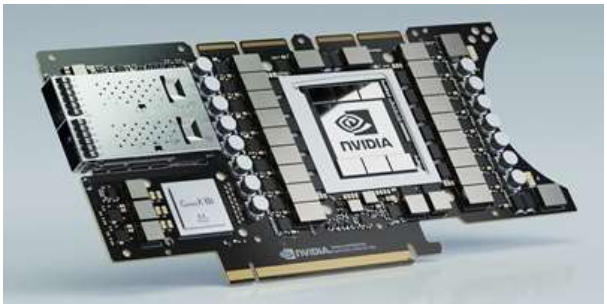
---

[449] Voir Nvidia GeForce RTX 30 Series & Ampere GPUs Further Detailed – GA102/GA104 GPU Specs & RTX 3090, RTX 3080, RTX 3070 Performance & Features Revealed, septembre 2020.

[450] Depuis avril 2019, Nvidia est concurrencé par l'annonce faite par Tesla de sa carte FSD qui intègre un processeur maison dédié à la conduite autonome. Il serait 7 fois plus rapide que la Nvidia Drive avec 144 teraflops vs 21 teraflops (malgré le fait que Nvidia communique en teraops pour la Drive).

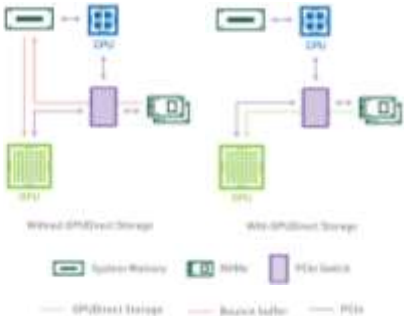
[451] Voir Nvidia Announces New Drive Platforms With Orin and Ampere par Andrei Frumusanu, mai 2020.

- Une déclinaison du A100 pour l'embarqué avec la carte **Jetson** EGX A100 qui a l'air de contenir un GPU A100 normal. Elle peut notamment servir à analyser les images provenant de dizaines de caméras de surveillance. Et par exemple, de compter la proportion des gens qui portent des masques dans les lieux publics surveillés ! Le covid revient par la fenêtre !



- **Jetson Xavier NX** est une petite carte de 70x45 mm plus petite qu'une carte de crédit. Avec 14 TOPs, 10W, ou 21 TOPS à 15W. Elle est jusqu'à 15 fois plus rapide qu'une Jetson TX2. Elle comprend un GPU Volta GPU avec 384 cœurs CUDA et 48 tenseurs, un CPU avec 6 cœurs arm 64 bits, 8 Go de mémoire.

En août 2019, Nvidia annonçait **GPUDirect Storage**, une nouvelle technologie permettant d'accélérer l'accès aux données stockées dans des SSD via l'interface NVMe par ses GPU[452].



Elle court-circuite le CPU pour les transferts de données, ce qui permet d'exécuter des applications de machine learning plus efficacement. Cela permet en gros de doubler le débit de lecture/écriture et de réduire la latence d'accès d'un facteur 3,8.

En octobre 2020, Nvidia lançait de nouvelles cartes réseaux pour serveurs (NIC - network adapter cards), des **DPU** (Data Processing Unit) de sa division Mellanox, la Bluefield-2 (*ci-dessous* à gauche) et la Bluefield-2X (*ci-dessous*, à droite), intégrant un processeur à 8 cœurs Arm et cette dernière intégrant un GPU A100. **VMware** qui facilite le déploiement des GPGPU Nvidia dans le cloud en les supportant dans VMware vSphere (virtualisation de serveurs), Cloud Foundation (plateforme de cloud hybride) et Tanzu (catalogue d'applications tierces qui supportent les containers Kubernetes)[453].

[452] Voir Nvidia GPU accelerators get a direct pipe to big data par Michael Feldman, août 2019.

[453] Voir Partenariat VMware / Nvidia sur l'IA dans le datacenter et le cloud par Lary Dignan, septembre 2020.

HPC

Les GPGPU de Nvidia se sont taillé la part du lion de l'équipement des supercalculateurs occidentaux depuis 2017. Le V100 équipe l'IBM Summit depuis 2018. Il est en fait devenu omniprésent. On n'y échappe guère qu'en Asie, en Chine ou au Japon (avec le Fugaku de Fujitsu).

Dernier en date, le supercalculateur de référence **DGX A100 SuperPOD** proposé par Nvidia comprend jusqu'à 280 serveurs DGX A100 avec 2240 GPGPU A100 et 7 Po de stockage. A raison de 4 DGX par rack, cela nécessite donc au minimum 70 racks.

Dénommé Selene, il a été installé en un mois pendant l'été 2020 dans la Silicon Valley[454]. Il est dédié au calcul scientifique dans la santé.

Une déclinaison du SuperPOD doit être déployée au Royaume-Uni sous l'appellation de **Cambridge-1** pour délivrer 400 PFLOPS (en application de machine learning) et 8 PFLOPS (pour le calcul scientifique, selon le benchmark Linpac) avec 80 serveurs DGX A100.

Il se retrouvera ainsi dans le TOP30 des plus grands supercalculateurs mondiaux[455].

Dans cette lignée, **Atos** annonçait en 2020 le déploiement pour le centre de calcul italien CINECA un supercalculateur de 200 PFLOPS. Ce Leonardo comprendra 14 000 GPU A100. Atos fournira aussi un supercalculateur au Royaume-Uni destiné à l'Université d'Oxford dans le cadre du projet JADE2. Il sera hébergé par le Hartree Center du STFC, l'équivalent du CEA, pour £5M. Exploitant l'architecture Nvidia DGX SuperPOD, il comprend 504 V100 dans 63 serveurs DGX, le tout relié comme il se doit par une connectique Mellanox InfiniBand et complété par un stockage DDN AI400 de **DataDirect Networks** (1998, USA, $9,9M). Enfin, Atos fournissait en septembre 2020 le supercalculateur Juwels aux laboratoires Jülich en Allemagne[456]. Il comprend des centaines de lames BullSequana X215 intégrant quatre GPGPU A100, délivrant un total 70 PFLOPS. Le refroidissement est réalisé par eau chaude, mais qui est plus froide que les composants à refroidir.

On peut aussi signaler dans cette lignée le supercalculateur que **Microsoft** déployait en mai 2020. Il comprend 10 000 GPU non précisés, dont on présume qu'il s'agit alors de Nvidia V100 au vu de la date du déploiement[457].

En 2019, le supercalculateur le plus puissant au monde était l'**IBM Summit** du centre de recherche d'Oak Ridge du Département de l'Energie dans le Tennessee.

---

[454] Voir AI of the Storm: How We Built the Most Powerful Industrial Computer in the U.S. in Three Weeks During a Pandemic, août 2020.

[455] Voir NVIDIA Will Build The UK's "Most Powerful" Supercomputer To Boost COVID-19 Research par Carly Page, octobre 2020.

[456] Voir Atos Launches First Supercomputer Equipped with NVIDIA A100 GPU, mai 2020.

[457] Voir Microsoft dévoile son superordinateur : 285 000 cœurs, 10 000 GPU et 400 Gbit/s, 2020 et Microsoft teamed up with OpenAI to build a massive AI supercomputer in Azure par Frederic Lardinois, mai 2020.

Il consomme 13 MW pour une puissance crête de 200 PFLOPS et dont 3,9 MW juste pour pour le refroidissement (source). Ces MW sont consommés par 202 752 CPU IBM Power9 et 27 648 GPU Nvidia V100. Le Summit occupe 500 m^2 et pèse 349 tonnes. Comme il est équipé de GPU Nvidia, il peut très bien servir à des applications d'IA. La quasi-totalité des supercalculateurs lancés depuis 2018 sont ainsi équipés de Nvidia V100 pour faire de l'entraînement de réseaux de neurones dans divers cas d'applications.

En France, le GENCI (qui associe le CNRS, le CEA et Inria) déployait au printemps 2019 son supercalculateur Jean Zay fourni par **HPE**[458]. Il était upgradé lors de l'été 2020 pour atteindre une puissance de calcul de 28 PFLOPS/s[459]. Il est équipé de CPUs Intel Cascade Lake 6248 et de 2696 GPU Nvidia V100[460], le tout associé à 1 Po de stockage en SSD. Il consomme un total de 2MW, occupe 150 m^2 au sol et pèse 43 tonnes. Il est installé dans le centre de calcul IDRIS à Orsay.

Fin 2019, **IBM** déployait ce qu'ils présentaient alors comme le plus grand supercalculateur installé dans une université privée, l'AiMOS (Artificial Intelligence Multiprocessing Optimized System) du Rensselaer Polytechnic Institute Center for Computational Innovations (CCI) de la SUNY Polytechnic Institute à New York[461]. Il est dédié à l'utilisation du machine learning pour optimiser la conception de chipsets d'IA. Sa configuration comprend 252 nœuds intégrant 504 CPU IBM Power9 et 1512 GPGPU Nvidia V100 similaires à ceux de l'IBM Summit (avec 2 CPU et 6 GPU par nœud. Il comprend un total de 128 Po de mémoire et délivre 8 PFLOPS (selon le benchmark Linpack utilisé pour le calcul scientifique). Le tout consomme 512 kW. C'est à peu près l'équivalent d'un demi Jean Zay du GENCI.

La course aux chipsets de l'IA a évolué ces dernières années. L'approche de Nvidia a illustré la nécessité pour ces fournisseurs de couvrir aussi bien les besoins du machine learning que ceux du calcul scientifique en général. Les chipsets d'IA dédiés aux serveurs sont donc conçus pour bien répondre à ces deux besoins, tout comme les bibliothèques logicielles associées[462].

Nvidia n'est cependant pas seul à déployer des GPGPU dans des supercalculateurs. **AMD** fait de même, à une échelle tout aussi impressionnante. Ses CPU et GPU sont ainsi en cours de déploiement dans deux supercalculateurs exascale de HPE/Cray au Département de l'Energie US (DoE). Il y a d'abord le Frontier qui doit être déployé à Oak Ridge en 2021 puis El Capitan au Lawrence Livemore lab en 2023. Les deux supercalculateurs exploitent des CPU AMD EPYC et des accélérateurs GPU Radeon Instinct sous forme de cartes PCIe, partageant l'accès à la même mémoire que les CPU.

---

[458] HPE s'offrait au passage le doyen indépendant des supercalculateurs, Cray, en avril 2019 pour $1,3B. HPE et Cray ont d'ailleurs été sélectionnés pour installer un supercalculateur au centre Argonne du DoE US en 2021 de capacité exascale (exaflops). Les Cray associaient déjà une combinaison de CPU Intel, de GPU Nvidia et de CPU Cavium à noyaux Arm. Voir France Merges Cascade Lake Xeons With Volta Tesla GPUs For AI Supercomputer par Michael Feldman, 2019. En 2020, plus de 300 applications différentes tournaient dans le Jean Zay, essentiellement autour du traitement du langage et de la vision artificielle. En voici quelques exemples : l'étude de la réionisation de l'Univers un milliard d'années après sa naissance, réalisée par l'Université de Strasbourg, l'analyse de fonctions du cerveau couplant de la simulation et de l'IRM fonctionnelle réalisée par le CEA Institut des Sciences du Vivant Joliot Curie et l'Université Paris Saclay, l'analyse d'IRM du cerveau pour détecter les neurodégénérescences liées à l'âge au laboratoire ARAMIS et l'étude d'attaques adversariales de réseaux de neurones réalisée par le laboratoire LAMSADE de l'Université Paris-Dauphine.

[459] Voir Le supercalculateur français Jean Zay atteint un record de 28 millions d'opérations par seconde par Benjamin Bruel, octobre 2020.

[460] Facebook a financé 200 des GPU V100 du Jean Zay.

[461] Voir IBM's latest supercomputer will be used... to build even more computers par Daphne Leprince-Ringuet, décembre 2019.

[462] Voir Another Strange Shift for the AI Chip Startup Segment par Nicole Hemsoth, septembre 2020 qui évoque les startups qui ont d'abord migré de l'entraînement et l'inférence de réseaux de neurones, et qui ciblent maintenant des applications de calcul plus générales intégrant le calcul scientifique, comme le fait Nvidia avec ses V100 et A100. Cela passe par le support de fonctions d'algèbre linéaire et le besoin de support logiciel généraliste. Cela permet notamment de couvrir le gros marché en volume des centres de calcul des laboratoires nationaux comme ceux du DoE aux USA ou du GENCI en France. Exemple de besoin en algèbre linéaire : la décomposition de matrices hilbertiennes en matrices triangulaires (LL^*, L^* étant la transconjugée de L), selon la méthode de décomposition de Cholesky.

Le Frontier fera 1,5 exaflops et l'El Capitan atteindra 2 exaflops, en consommant 40 MW[463]. Ces calculateurs serviront surtout à faire de la simulation 3D et notamment à simuler l'arme atomique US. Mais ils auront aussi des usages dans le machine learning.

CUDA

Il nous faut maintenant rappeler l'importance de l'architecture logicielle CUDA qui pilote tous ces processeurs[464]. C'est une API de programmation à bas niveau des GPU de Nvidia. Elle est supportée par diverses bibliothèques fonctionnelles proposées par Nvidia aussi bien pour la gestion de réseaux de neurones que pour le calcul scientifique. Les APIs de CUDA sont par ailleurs supportées et générées par les principaux framework de machine learning et de deep learning du marché à commencer par TensorFlow et PyTorch.

La sortie en mai 2020 du GPGPU A100 était accompagnée d'une nouvelle version 11 de la bibliothèque logicielle CUDA. Nvidia lançait aussi une nouvelle version de son environnement de développement Nsight qui supporte notamment le profiling, le débogage et l'analyse de code ciblant le A100 ainsi que les noyaux arm et Power (d'IBM). Avec l'acquisition d'arm par Nvidia en octobre 2020, on peut s'attendre à ce que ce dernier impose son API CUDA dans les systèmes embarqués où dominent les noyaux arm.

L'offre logicielle de Nvidia et son écosystème de développeurs font de son offre de processeurs une véritable plateforme complète. Cela génère une significative barrière à l'entrée pour les startups qui se lancent dans la création de chipsets dédiés au machine learning. Cette barrière à l'entrée est à la fois technique et marketing. On le retrouve ainsi du côté des benchmarks où très peu de startups arrivent à faire significativement mieux que Nvidia[465].

Processeurs neuromorphiques

Les processeurs neuromorphiques auraient été conceptualisés pour la première fois en 1990, par **Carver Mead**[466]. Ils sont spécialisés dans les traitements de machine learning et deep learning. Ils exécutent les processus d'entraînement puis d'exécution d'applications de deep learning et leurs réseaux de neurones. Il existe cependant de nombreuses variantes de tels processeurs neuromorphiques tant au niveau logique que physique[467]. Les variantes se situent dans l'intégration de fonctions de calcul vectoriel, matriciel, des matrices de neurones interconnectées, des neurones à impulsion et des approches hybrides intégrant plusieurs de ces techniques. Des chercheurs planchent aussi sur l'utilisation de composants fonctionnant à température supraconductrice[468].

Certains d'entre eux sont plutôt adaptés à l'entraînement de réseaux convolutifs avec des multiplicateurs de matrices et d'autres semblent conçus plutôt pour les réseaux à mémoire et le traitement du langage.

---

[463] Voir Cray to Build El Capitan Exascale Supercomputer at LLNL, août 2019 et El Capitan Supercomputer Detailed: AMD CPUs & GPUs To Drive 2 Exaflops of Compute par Ryan Smith, mars 2020.

[464] Voir Lecture 1: an introduction to CUDA par Mike Giles (247 slides) qui décrit les principes généraux de la programmation en CUDA, Tensor Core Programmability and Profiling for AI and HPC applications par Griffin Lacey et Max Katz, 2019 (47 slides) qui couvre les besoins aussi bien côté IA que calcul scientifique, CUDA new features and beyond par Stephen Jones, 2019 (73 slides), NVIDIA Tensor Core Programmability, Performance & Precision par Stefano Markidis et al, 2018 (12 pages), Introduction to GPU computing with CUDA par Pierre Kestener, 2016 (241 slides) qui décrit bien l'histoire des CPU et des GPU et les gains de performance associés.

[465] Voir Why Can't NVIDIA Be Bested In MLPerf? par Karl Freund, octobre 2020.

[466] Voir Neuromorphic Electronic Systems, Carver Mead, 1990.

[467] Voir cet inventaire de différents types d'architectures de processeurs neuromorphiques : Neuromorphic Computing and Neural Network Hardware, 2017 (88 pages) qui s'appuie sur une bibliographie record de 2682 entrées. Vous pouvez aussi consulter l'excellente présentation Hardware Architectures for Deep Neural Networks, 2017 (290 slides).

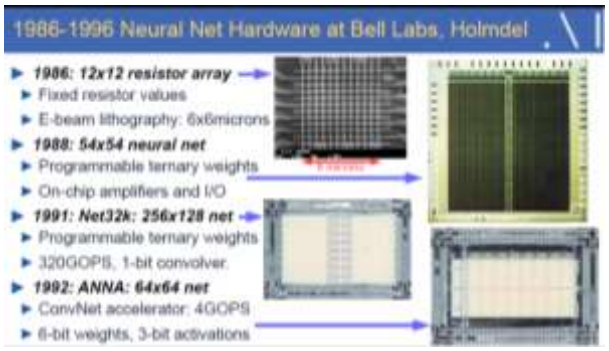
[468] Voir A Power Efficient Artificial Neuron Using Superconducting Nanowires par Emily Toomey, Ken Segall et Karl Berggren, 2019 (17 pages).

Cela se traduit par des architectures physiques différentes qui sont plus ou moins proches de l'organisation de neurones physiques mises à plat dans leur layout[469].

Les chipsets adaptés aux réseaux convolutifs comprennent des unités de multiplication de matrices et de vecteurs qui sont utilisées dans les premières couches de convolution (NPU ou TPU, neural processing units ou tensor processing units). Les dernières couches « fully connected » s'appuient aussi sur des matrices reliant les neurones d'entrées avec les neurones des couches suivantes, via des grilles comprenant les poids des synapses. Mathématiquement, ce sont des unités de traitement qui multiplient des vecteurs à une dimension par des matrices pour générer des vecteurs, mais avec des fonctions d'activation généralement non linéaires au niveau de chaque neurone.

Ces chipsets sont conçus pour que la mémoire qui stocke les poids des synapses et les *feature maps* des réseaux convolutifs soit la plus proche des unités de traitement, afin d'en accélérer le fonctionnement, surtout pendant les phases d'entraînement[470], le cas le plus avancé étant celui des memristors.

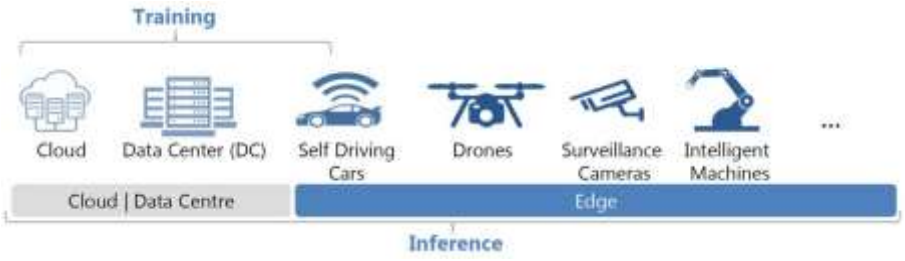
Les premiers chipsets pour réseaux de neurones sont en fait apparus entre 1986 et 1992, avec une puissance de quelques centaines de GOPS (giga opérations par seconde) notamment dans les **Bell Labs** comme le raconte bien Yann Le Cun. Mais il s'agissait à l'époque d'objets de laboratoires, pas de produits commercialisés en volume[471].



L'autre variante se situe dans la position des chipsets dans la chaîne de valeur de l'IA.

Certains chipsets très puissants sont adaptés à l'entraînement de réseaux de neurones sur des serveurs tandis que d'autres sont dédiés à l'exécution ou « inférences » de réseaux de neurones déjà entraînés dans des objets connectés ou smartphones.

L'un des enjeux est pouvoir à terme faire de l'apprentissage par renforcement dans des chipsets embarqués tout en préservant une basse consommation d'énergie.



Ces deux dimensions sont reflétées dans le schéma suivant de mon cru avec une bonne part des startups et sociétés citées dans ce qui va suivre. On y voit que l'offre est bien plus abondante du côté de l'embarqué que des serveurs. Le segment de l'edge computing correspond aux calculs réalisés près des objets, par exemple dans une centrale de contrôle de caméras de surveillance. Un end point est un objet dans lequel on peut aussi intégrer des composants neuromorphiques.

Cette abondance provient notamment du fait que les composants côté embarqué sont souvent plus spécialisés alors que les chipsets pour serveurs sont plutôt génériques et coûtent plus cher à concevoir et à fabriquer.

---

[469] Voir quelques exemples de neurones plus ou moins proches des neurones biologiques : Brain synapses : Ultralow power artificial synapses using nanotextured magnetic Josephson junctions, janvier 2018 et Artificial Brain Synapses Replicated in a Chip, février 2018.

[470] C'est une approche qui est aussi adoptée par la startup grenobloise **UpMem** (2015, France, $3,6M) qui conçoit des circuits de traitement intégrant mémoire et calcul (Processing-In-Memory ou PIM), mais dédiés au big data. Visiblement, l'architecture semble plus proche de celle des GPU que des processeurs neuromorphiques.

[471] Source du schéma : The Power and Limits Of Deep Learning, Yann Le Cun, conférence à Harvard, mars 2019.

Cela va probablement déboucher sur une consolidation du marché, du fait des économies d'échelle et des effets de plateformisation des offres. Les fournisseurs de chipsets, surtout d'ASIC, ne peuvent pas se contenter de cibler des marchés de niche sans économies d'échelle[472].

Ce marché est pour l'instant très fragmenté avec une offre pléthorique, pas toujours bien différenciée. Le nombre d'acteurs chinois y est important avec des acteurs bien financés, mais qui n'ont pas encore l'influence sur le marché occidental qu'ont des acteurs tels que Nvidia[473].

- **ASIC** : ce sont des chipsets fabriqués en volume dont l'agencement (layout) est défini une fois pour toute avant la fonderie. C'est la technique utilisée pour GPU de Nvidia, pour les processeurs Intel et pour les processeurs de smartphones. Elle est adaptée aux gros volumes de production. Elle est aussi plus efficace côté puissance et économie d'énergie par rapport aux FPGA, pouvant aller jusqu'à un rapport de 1 pour 100 à 1000. C'est l'approche retenue par Google pour ses TPU[474].

On peut aussi classifier ces processeurs suivant un autre axe lié à leur mode de réalisation physique :

- **FPGA** : ce sont des processeurs programmables qui sont généralement créés pour de faibles volumes de production[475]. Ils servent surtout à exécuter des modèles de machine learning et de deep learning déjà entraînés[476]. On peut activer par logiciel les portes logiques du circuit pour gérer des modèles sur mesure. Ils sont un peu l'équivalent de l'impression 3D pour les chipsets : intéressants pour les faibles volumes et le prototypage rapide mais moins pour les volumes importants.

---

[472] Voir Getting Design Wins for AI Accelerators par Andrew Richards, novembre 2020, qui décrit les approches d'intégration horizontale ou verticale du marché des chipsets d'IA. Et cartographie de chipsets d'IA : https://basicmi.github.io/AI-Chip/.

[473] Voir Competing in Artificial Intelligence Chips China's Challenge amid Technology War par Dieter Ernst 2020 (70 pages) qui décrit la nature de cette concurrence chinoise.

[474] L'approche de Google est décrite en détails dans First in-depth look at Google's TPU architecture, par Nicole Hemsoth, avril 2017. Cela concerne leur première génération de TPU. Les suivantes ont conservé une architecture similaire, en diminuant la taille des matrices gérées.

[475] Voir FPGA-based Accelerators of Deep Learning Networks for Learning and Classification: A Review par Ahmad Shawahna, 2019 (41 pages).

[476] Des travaux visent à utiliser les FPGA pour réaliser de l'entraînement de réseaux de neurones et de manière efficace énergétiquement. Voir Challenges in Energy-Efficient Deep Neural Network Training with FPGA par Yudong Tao et al, 2020 (10 pages).

C'est la technologie retenue par **Microsoft** pour ses chipsets Brainwave qui exploite des FPGA d'origine Intel, en plus des inévitables GPU de Nvidia. On en trouve aussi chez diverses startups comme **Teradeep** (2014, USA) ou **Leapmind** (2012, Japon, $13,4M). Ces processeurs peuvent être 10 fois plus rapides que des GPU.

La dernière génération des puces FPGA Versal **Xilinx** gravée en 7 nm combine maintenant des cœurs arm, des DSP de traitement du signal audio et vidéo et de réseaux de neurones et une partie programmable, apportant une souplesse et une performance utile à de nombreux cas d'usages[477]. Xilink fournit aussi des cartes d'accélération Alveo plus classiques adaptées aux inférences avec jusqu'à un million de LUT (Look-Up Tables) qui sont les unités logiques de base de calcul des FPGA[478]. Notons qu'en octobre 2020, AMD faisait une offre d'acquisition de Xilinx pour $35B. La consolidation du marché se poursuit, Intel ayant acquis Altera en 2015.

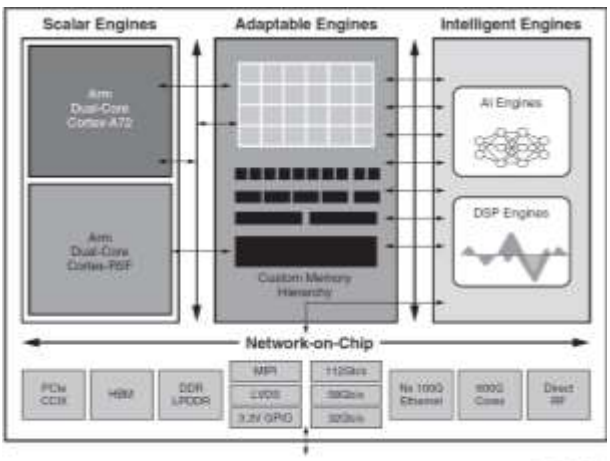
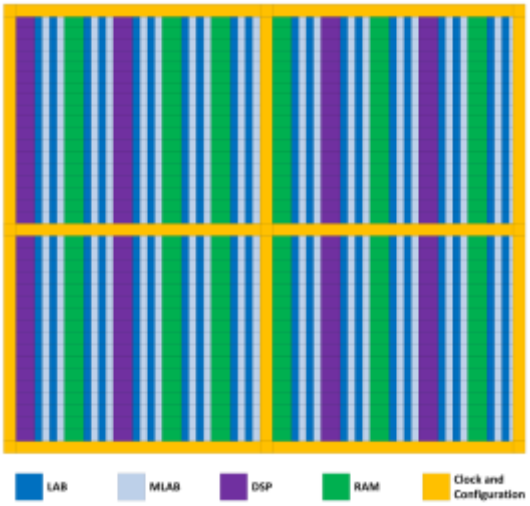


Figure 4: **Xilinx Versal ACAP Functional Diagram**



Chez **Intel**, la dernière génération avancée de FPGA est l'Agilex gravée en 10 nm. Elle contient aussi des noyaux arm, quatre cœurs 64-bits Cortex A53, et optimise le routage et l'accès à la mémoire qui est réparti dans le circuit. Les Agilex supportent aussi des entrées/sorties optimisées (4 fois 400 Gbits/s Ethernet) dont le bus PCIe 4 et 5, le tout supportant 40 TFLOPS. Dans la gamme des FPGA Stratix gravée en 14 nm, le Stratix NX lancé en 2020 intègre aussi des tenseurs pour accélérer le calcul matriciel et une liaison rapide avec de la mémoire externe HBM2 pouvant atteindre 512 Go/s par puce[479].

L'une des manières d'améliorer la performance des FPGA est l'utilisation de la technique du NoC (Network on Chip) qui relie les unités de calcul programmable par des liaisons réseaux maillées à très haute vitesse (jusqu'à 512 Gbits/s), ce que propose **Achronix** (USA) avec ses FPGA Speedster7t.

- **Memristors** : ce sont des circuits de réseaux de neurones qui mémorisent de manière non volatile les poids des synapses des neurones au sein même des neurones. Dans les FPGA ou ASIC, ces informations sont stockées dans des puces mémoires séparées et de manière volatile. Le stockage de ce poids des neurones évite des accès mémoire externes et permet une accélération significative de la phase d'entraînement. Ces poids sont gérés sous forme de résistances variables non volatiles et donc sous forme analogique.

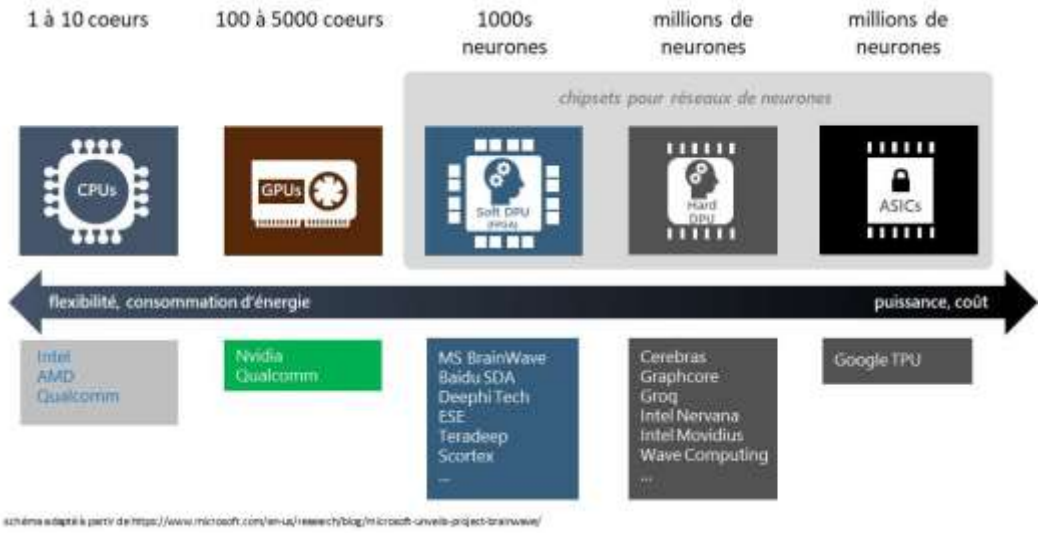
---

[477] Voir Versal: The First Adaptive Compute Acceleration Platform (ACAP), 2020 (21 pages) et Versal Premium Series Announcement, Mike Thompson, 2020 (34 slides). Et aussi une description de l'offre Xilinx dans Powering Cloud and Datacenters with Xilinx Adaptive Compute platforms par Cathal McCabe, mai 2020 (36 slides).

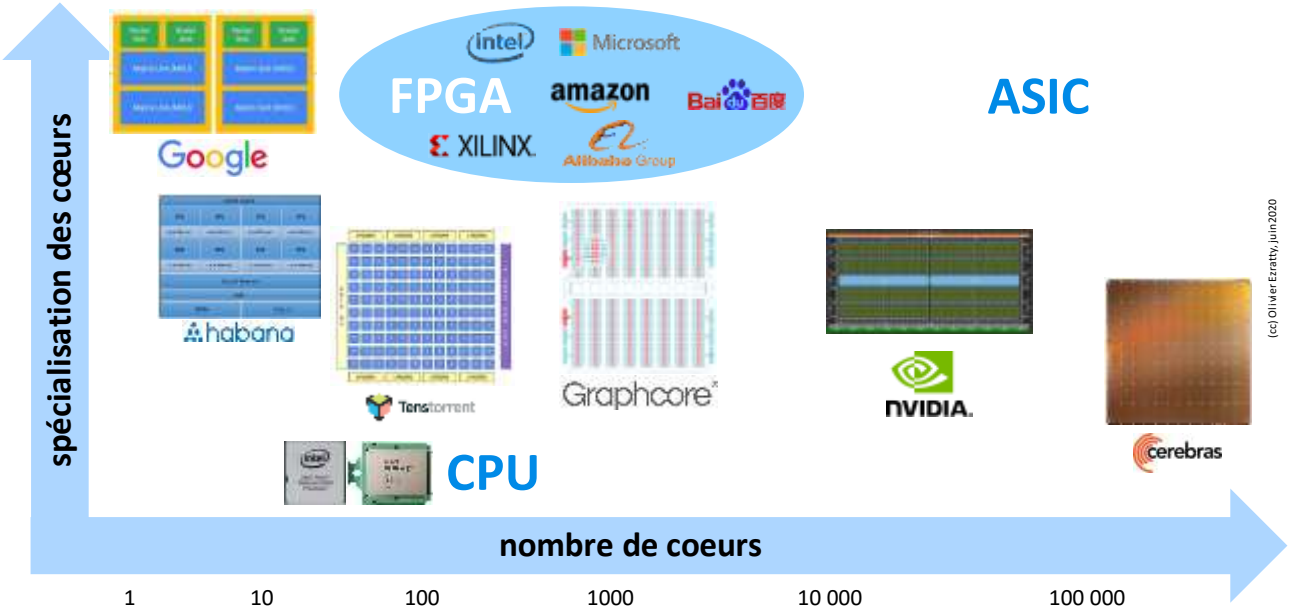
[478] Les LookUp Tables gèrent des tables de correspondance entre n bits en entrée et un bit en sortie. Ce sont en quelque sorte des portes logiques génériques et paramétrales dynamiquement.

[479] Voir HBM2, Tensor Units Key to Intel's New AI Centric FPGA par Nicole Hemsworth, novembre 2020.

Cela présente aussi l'avantage d'accélérer les phases d'entraînement par rétropropagation d'erreurs dans les réseaux de neurones.



Comme certains acteurs tels qu'Intel sont présents sur plusieurs créneaux à la fois, je vais ici décrire leur offre, fournisseur par fournisseur.



Il est probable qu'après la lecture de ces plus de 90 pages sur les processeurs, la mémoire et le stockage, vous serez un peu perdus et vous demanderez : comment vais-je devoir m'équiper pour ma solution de deep learning, et comment vais-je répartir les traitements entre le cloud et les objets. Ne vous inquiétez pas, vous n'êtes pas seul [480]! Le savoir sur le sujet est en pleine évolution. Les capacités des processeurs évoluent sans cesse. Il faut donc naviguer à vue mais en comprenant les tenants et aboutissants des différentes technologies proposées, ce que j'essaye de faire ici.

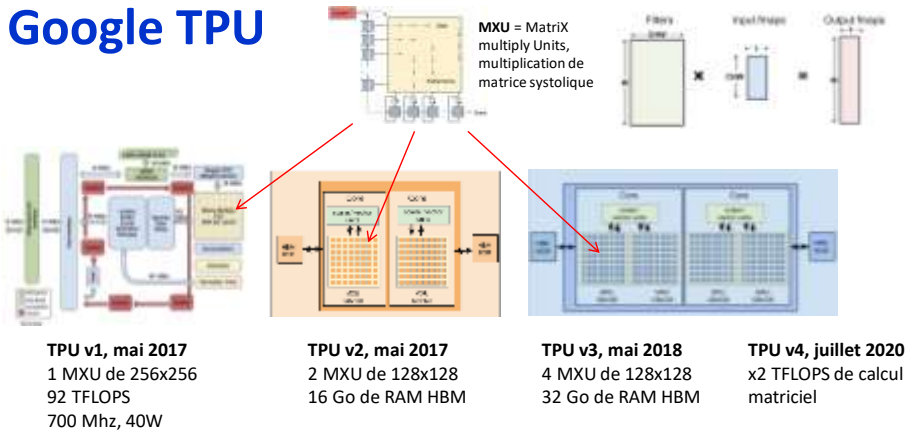
Google

Google a créé ses TPU ou Tensor Processing Units adaptés notamment à l'exécution des applications développées avec TensorFlow.

[480] Voir Why AI adoption leaves customers dazed and confused, par Michael Feldman, avril 2019.

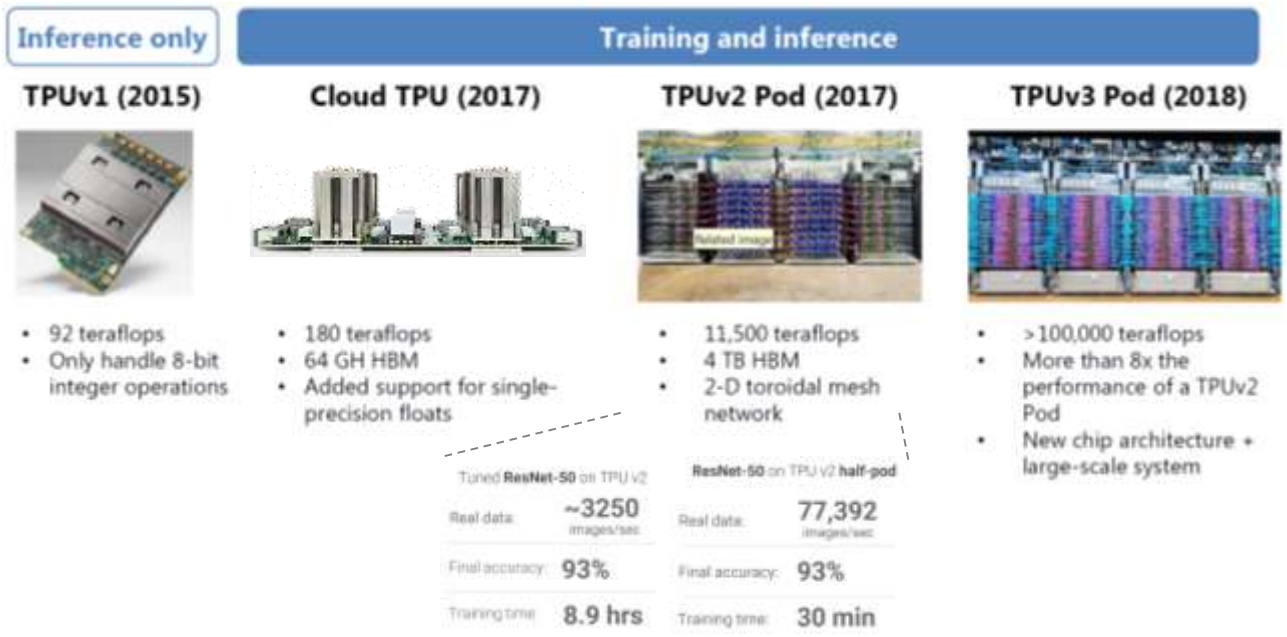
Ce sont eux qui ont permis la victoire d'AlphaGo au jeu de Go début 2016. Ils sont intégrés dans les datacenters de Google pour ses applications et services en cloud mais ne sont pas commercialisés séparément. Ils en étaient à leur quatrième génération de TPU en juillet 2020. Ils utilisent des ASIC performants et consommant peu d'énergie.

Leur layout semble surtout adapté à l'exécution de réseaux convolutifs avec une matrice de 128x128 pour gérer des convolutions ainsi que des couches de pooling (réduction de résolution) et d'activation (dernière couche de réseaux de neurones pour obtenir les tags des objets détectés)[481].



Les TPU v3 étaient huit fois plus puissants que les v2, sans que l'on dispose de plus de détails techniques autrement que leur capacité mémoire de 32 Go au standard HBM et le fait qu'ils doivent être refroidis par liquide[482]. Ils sont passés de matrices 256x256 à 128x128 pour gérer les convolutions. C'est le rôle des compilateurs de convertir les matrices des algorithmes de deep learning en matrices de la taille gérée par le processeur[483].

Les spécifications et le layout fonctionnel du TPU v4 n'étaient pas encore disponibles en décembre 2020. On sait juste qu'ils doublent la capacité de calcul matriciel en TFLOPS par rapport aux TPU v3.



[481] Voir des détails sur l'architecture des TPU sur A Domain-Specific Architecture for Deep Neural Networks, septembre 2018.

[482] Comme les GPU de Nvidia, les TPU de Google peuvent aussi servir à autre chose qu'au machine learning. Ils peuvent notamment servir à retraiter des images médicales via des transformées de Fourier discrètes. Voir Large-Scale Discrete Fourier Transform on TPUs par Tianjian Lu, 2020 (10 pages).

[483] Voir Cloud TPU: Codesigning Architecture and Infrastructure par Clifford Chao, août 2019 (64 slides) qui évoque les systolic arrays utilisées dans les TPU Google et les GEMM (general matrix multiply).

Les TPU désignent aussi bien les racks de serveurs que les chipsets qu'ils contiennent[484]. Depuis 2017, les TPU servent à l'entraînement de réseaux de neurones ainsi qu'à leurs inférences alors que la première version de 2015 ne gérait que les inférences.

Les serveurs de TPU de Google son arrangés en Pods, de grandes rangées de racks de serveurs, dont les versions ne suivent pas forcément celle des ASIC TPU qu'ils contiennent. Ainsi, les TPUv4 Pod de 2020 contenaient-ils 4096 chipsets TPU v3 totalisant 430 PFLOPS.

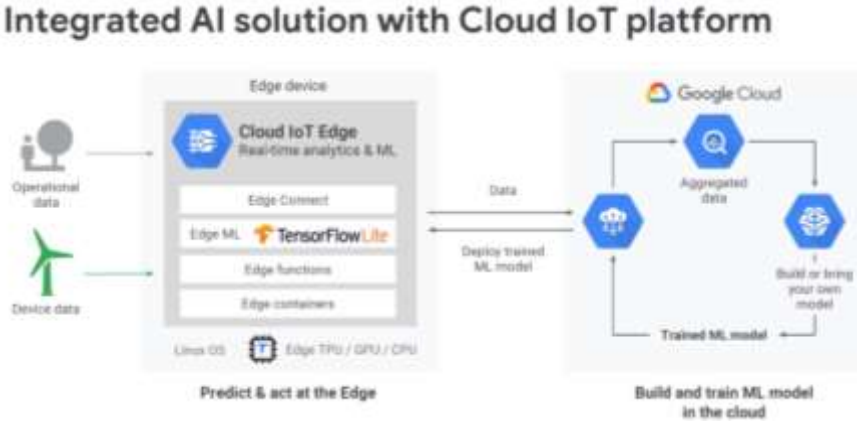
En juillet 2018, Google complétait le tableau avec ses cartes **Edge TPU**, des versions adaptées aux objets connectés, sans que les spécifications détaillées ne soient vraiment disponibles pour la puce TPU ssociée.

Edge TPU, 2019

Coral Edge TPU, 2020

L'offre comprend une carte de développement équipée de chipsets à noyaux ARM et Vivante pour le GPU (Edge TPU Dev Board) et une carte USB équipée d'un TPU Google.

Tous les deux sont adaptés à l'exécution d'inférences de réseaux de neurones développés avec TensorFlow Lite[485].



TensorFlow Lite est même décliné pour fonctionner sur des micro-contrôleurs à basse consommation[486].

Du coté des smartphones, Google avait développé son propre composant de machine learning pour le traitement des images pour ses smartphones Pixel avec le **Pixel Visual Core** intégré dans les Pixel 2 et 3 puis le Pixel Neural Core dans les Pixel 4. Google a abandonné cette approche dans le Pixel 5 qui exploite les fonctions de traitement de l'image du SoC Qualcomm Snapdragon 765G.

Notons enfin que Google utilise aussi le machine learning pour optimiser le routage dans ses circuits[487].

Intel

Intel a bien du mal à se distinguer dans ce nouveau paysage des chipsets plus ou moins dédiés au machine learning. Leur offre n'est pas très lisible. Elle couvre l'entraînement et l'exécution d'applications de deep learning des serveurs aux objets connectés, une bonne part provenant d'acquisitions plus ou moins bien digérées.

---

[484] Voir cette description assez détaillée sur Tearing apart Google's TPU 3.0 AI coprocessor, mai 2018. Le schéma sur les TPU est issu de la presentation AI optimized chipsets Part III de Vertex, 2018 (24 slides).

[485] Voir Intelligently bridging on-prem & Cloud with Edge TPU and Cloud IoT Edge 2018 (31 slides). Source du schema : Machine Learning on Google Cloud de Melanie Warrick, Google 2018 (80 slides).

[486] Voir Coming soon – TensorFlow Lite for Microcontrollers Kit, juillet 2019.

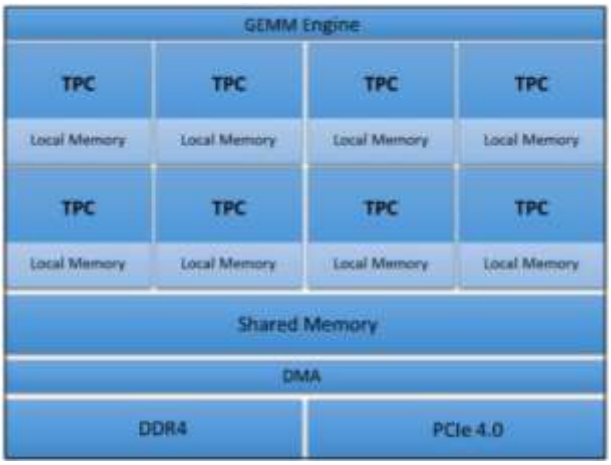
[487] Voir Google Invents AI That Learns a Key Part of Chip Design - AI helps designs AI chip that might help an AI design future AI chips par Samuel K. Moore, mars 2020, qui fait référence à Placement Optimization with Deep Reinforcement Learning par Anna Goldie et Azalia Mirhoseini, 2020 (5 pages).

Côté serveur, nous avons :

- **Xeon** et **Xeon Phi**, la gamme de processeurs généralistes dédiés aux serveurs. Intel fait des efforts pour optimiser les frameworks de deep learning (TensorFlow, Torch, Caffe) pour qu'ils s'exécutent plus rapidement sur des architectures Core et Xeon traditionnelles, alors qu'ils sont habituellement plutôt optimisés pour les GPUs de Nvidia. Cela aurait permis d'améliorer les performances d'un facteur x70 à x85 sur les processeurs Xeon[488] qui équipent les serveurs de data centers, rapprochant leurs performances des meilleurs GPU Nvidia de 2017. La mémoire adressable par ces chipsets permet d'entraîner des modèles de reconnaissance d'images assez grands et de traiter des images de 1024x1024 pixels dans des réseaux convolutifs ce qui est particulièrement intéressant dans l'imagerie médicale. Les dernières générations en date de Xeon (fin 2020) sont les Cooper Lake et Ice Lake qui succèdent aux Cascade Lake qui améliorent la performance en calculs en nombres flottants, notamment dans le format hybride bfloat16 (inspiré par Google et adopté par Nvidia). Les Cooper Lake sont adaptés aux serveurs intégrant 4 ou 8 chipsets et Ice Lake pour les machines à un ou deux chipsets[489].

- L'offre de chipsets dédiés au machine learning s'articule maintenant autour de l'offre de la startup **Habana** (2016, Israël, $120M) acquise en décembre 2019 pour $2B. Celle-ci comprend une carte PCIe pour serveurs comprenant leur processeur Goya HL-1000 qui peut traiter 15 000 images/seconde avec seulement 100W contre 3 211 images pour un Nvidia V100 et 320W.

  Le tout grâce à l'intégration d'un multiplicateur de matrices, le GEMM (General Matrix et Matrix Multiplication) qui s'appuie sur huit cœurs tensoriels exploitant leur mémoire locale et qui supportent le calcul sur nombres flottants et entiers entre 8 et 32 bits. Le système supporte TensorFlow et le format d'échange ONNX.

  Après cette acquisition, Intel poubellisait toute l'offre de chipsets spécialisés **Nervana** (2014, Israël, $24,4M) résultant d'une autre acquisition en 2014.

  Celle-ci, dénommée Crest comprenait des ASIC dédiés à l'entraînement (NNP-T 1000, 24 TPC) et à l'exécution (NNP-I 1000) de réseaux de neurones sur serveurs[490]! Le NNP-L 1000 Spring Crest annoncé en avril 2019 était destiné aux data centers et pour aussi bien de l'entraînement que de l'inférence.

- Intel annonçait en novembre 2019 le chipset **Ponte Vecchio**, un GPU censé concurrencer les GPGPU de Nvidia, dit de X[e] génération (qui arrivent après les générations 9, 10 et 11, en toute logique...). Devant être gravé en 7 nm, le produit a pris du retard, ne serait-ce que parce qu'Intel n'arrive pas à mettre au point la gravure en 7 nm et doit s'en remettre à TSMC. Ponte Vecchio comprend 35 milliards de transistors.

---

[488] Voir TensorFlow Optimizations on Modern Intel Architecture, août 2017 et New Optimizations Improve Deep Learning Frameworks For CPUs, octobre 2017. Ces optimisations s'appuient sur l'utilisation des instructions de traitements de vecteurs AVX2 des Xeon et AVX512 des Xeon Phi, ainsi que sur les versions 2017 des bibliothèques Intel Math Kernel Library (Intel MKL) et Intel Data Analytics Acceleration Library (Intel DAAL). Le jeu d'instruction AVX512 permet de réaliser des opérations matricielles voisines de celles des cœurs Tensor des TPU de Google et des GPU Nvidia GV100.

[489] A noter que la Chine s'est lancée dans la conception d'une variante du Xeon d'Intel. Ce *Jintide* a en fait été co-développé avec Intel. Il comprend des éléments de sécurité chinois qui remplace le TPM (Trusted Platform Module) d'Intel. Voir Surprise : en pleine guerre technologique USA-Chine, un processeur Xeon chinois apparaît par 01net, mai 2020.

[490] L'architecture des processeurs Crest est très bien décrite dans Intel focuses on scale out AI traning with new chip par Nicole Hemsoth, août 2019.

Son architecture associe des modules de traitement capables de fonctionner en mode SIMD (single instruction, multiple data, voisins des cœurs de CPU classiques) et en mode SIMT (single instruction, multiple threads, voisins des unités de traitement parallèles des GPU). Le chipset est organisé en 8 sub-slices qui équivalent aux streaming multiprocessor de Nvidia avec leur mémoire cache L1 et L2.

Chaque sub-slice comprend 8 execution units, elles-mêmes comprenant 7 threads capables d'exécuter des calculs sur 128 bits avec quatre instructions par cycle d'horloge. C'est dans ces execution units que le choix est fait de fonctionner en mode SIMD ou SIMT.

Les unités de calcul sont reliées à la mémoire HBM externe via des canaux mémoire rapides XEMF (X^E Memory Fabric) associés à un cache dénommé Rambo.

Ce chipset doit équiper le supercalculateur Aurora du Département de l'Energie US qui doit entrer en service en 2021[491]. On espère que Ponte Vecchio sera prêt à temps ! Mais le retard de livraison de Ponte Vecchio met en danger ce projet [492]!

Aurora sera un calculateur exascale, donc capable d'atteindre un exaflops de puissance de calcul. Les GPU Ponte Vecchio complètent des CPU Xeon d'architecture Saphire Rapids, gravée en 10 nm. Aurora est fourni par Cray, une filiale d'HP Enterprises. Aurora aura un total de 10 Po de mémoire vive, un stockage orienté objet distribué de plus de 230 Po avec une bande passante de 25 To/s. Les nœuds de calcul comprennent 2 Xeon et 6 GPU, partageant la même mémoire et reliés entre eux par une connectique réseau maison Cray, le « Cray Slingshot Network » supportant des liaisons à 200 Gbits/s similaires à ce que permettent les interconnexions Infiniband Mellanox exploitées par Nvidia avec ses GPU et les CPU Intel ou AMD. L'architecture Cray utilise des switches à 64 ports supportant une bande passante de 25,6 Tbits/s organisés dans une topologie minimisant les nœuds de passage entre serveurs et limitant les congestions de trafic. Ces switches sont équipés des chipsets Rosetta maison, des ASIC gravés en 16 nm chez TSMC et consommant jusqu'à 250 W.

- Intel propose aussi une gamme de FPGA issue de l'acquisition d'Altera. **Stratix 10** est une gamme fabriquée en technologie 14 nm[493]. Ils sont programmés par les clients d'Intel comme Microsoft le fait pour ses chipsets neuromorphiques BrainWave. Ils présentent l'avantage d'intégrer des blocs de mémoire de 20 Kbits, utiles pour l'accélération de l'entraînement ou l'inférence de réseaux de neurones. Les Stratix 10 tournent en théorie jusqu'à 1 GHz, avec une capacité de 9,8 Gflops et une puissance de 80 Gflops par Watt. Intel fournit son SDK OpenVino pour le développement d'applications de deep learning, notamment dans la vision.

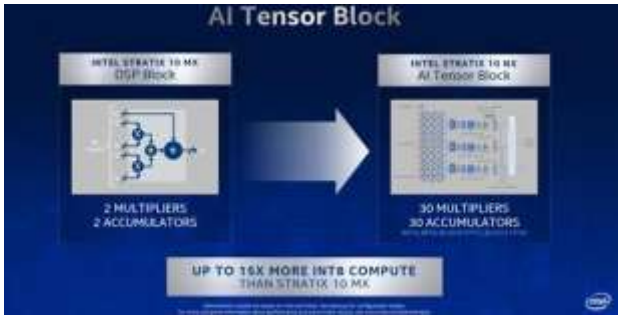
[491] Voir An Overview of Aurora, Argonne's Upcoming Exascale System par Yasaman Ghadar et Tim Williams, février 2020 (67 slides) et Inside Rosetta: The Engine Behind Cray's Slingshot Exascale-Era Interconnect, février 2020.

[492] Voir Exascale Exasperation: Why DOE Gave Intel a 2nd Chance; Can Nvidia GPUs Ride to Aurora's Rescue? par Doug Black, août 2020. Le chipset pourrait d'être disponible que début 2020, retardant d'au moins un an la mise en route du supercalculateur Aurora. En novembre 2020, Intel annonçait la disponibilité de cartes graphiques PCIe comprenant des chipsets X^e mais ce ne sont pas les versions devant équiper le supercalculateur Aurora. Voir Intel's First Discrete Xe Server GPU Aimed At Hyperscalers par Timothy Prickett Morgan, novembre 2020.

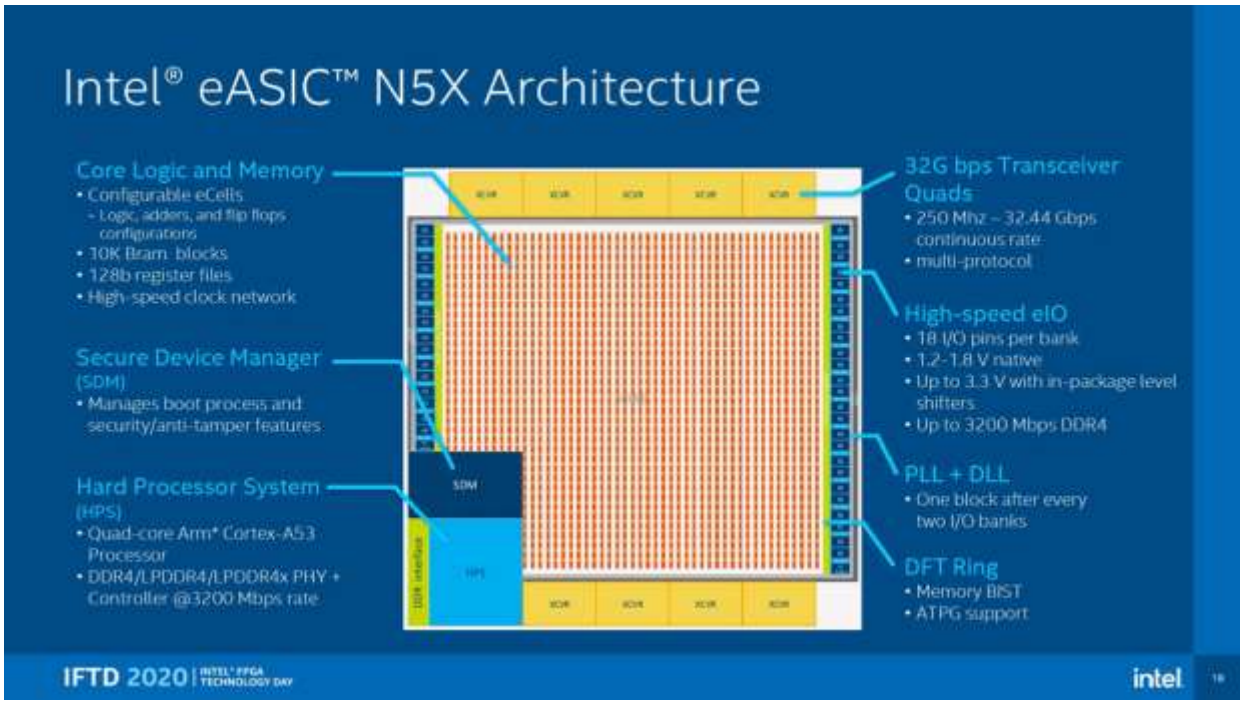
[493] Voir Machine Learning with Intel® FPGAs d'Adrian Macias, mai 2018 (32 slides).

Les chercheurs d'Intel militent d'ailleurs pour l'hybridation de solutions de deep learning intégrant des FPGA et des ASIC[494].

En avril 2019, Intel complétait son offre en faisant l'acquisition d'**Omnitek** (1998, UK), un fournisseur de blocs d'IP pour FPGA pour le traitement de l'image et de l'audio. La gamme Stratix a été ensuite complétée en 2019 par les **Agilex**, gravés en 10 nm. Lancés en juillet 2020, les Stratix 10 NX améliorent les performances pour le machine learning avec des blocs de calcul tensoriels intégrant jusqu'à 30 MACs (multipliers/accumulators) travaillant en int4, int8, int12 et int16, le support de mémoire HBM2 et une interconnectivité réseau haut débit.



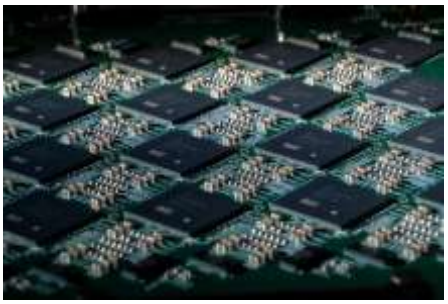
Cette gamme de FPGA est complétée par ce qu'Intel dénomme des eASIC ou structured ASIC, qui sont en fait des FPGA contenant des blocs d'IP ASIC fixes (DSP, tenseurs, sécurité, cœurs de CPU), notamment dans la série N5X lancée fin 2020. Ces eASIC sont issus de l'acquisition par Intel de la startup éponyme eASIC en 2018.

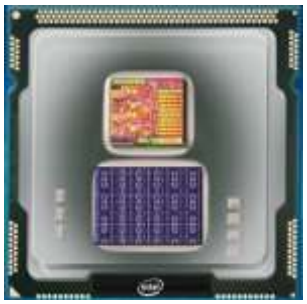


[494] Voir Why Compete When You Can Work Together: FPGA-ASIC Integration for Persistent RNNs par Eriko Nurvitadhi & Al, 2019 (9 pages). Cela s'applique ici à des réseaux de neurones récurrents (RNN). Voir aussi In-Package Domain-Specific ASICs for Intel® Stratix® 10 FPGAs: A Case Study of Accelerating Deep Learning Using TensorTile ASIC par Eriko Nurvitadhi & Al, 2018 (5 pages).

Pour ce qui est des inférences et de l'embarqué :

- **Movidius Myriad X VPU** est une gamme de processeurs vectoriels dédiés au traitement de l'image dans l'embarqué comme dans les caméras de surveillance ou dans les transports autonomes. Ces « Vision Processing Unit » permettent de traiter un trillion d'opérations de réseaux de neurones par secondes grâce à 16 processeurs vectoriels. Ils sont gravés en 16 nm chez TSMC. Le chipset peut aussi faire de l'encodage vidéo en 4K et ne consomme que 2W, à raison de 4 TOPS par watt. Tout cela est issu de l'acquisition de la startup Fathom en 2016. Le chipset Movidius Myriad 2 équipe les caméras infrarouges **FLIR** Firefly annoncées en octobre 2018 ainsi que les clés **USB Neural Compute Stick 2** qui peuvent être intégrées dans des serveurs, des PC ou des objets connectés.
- **MobileEye** est une filiale d'Intel, issue de son acquisition en 2017 pour $16B, qui a mis sur le marché depuis longtemps son propre chipset embarqué pour la reconnaissance d'image pour les véhicules à conduite assistée et autonome exploitant une ou plusieurs caméras RGB classiques.
- **GNA (Gaussian Network Accelerator)** est dédié au traitement du langage et à la vision. C'est un coprocesseur destiné aux processeurs des séries Atom et Core. Il semblerait que ses fonctions soient directement intégrées dans les chipsets Core de la 10e génération Intel lancés en 2019 et fabriqués en 10 nm.
- **Loihi** est une gamme de chipsets de neurones à impulsions annoncée en septembre 2017. Le chipset comprend 130 000 neurones à impulsions, comme dans les chipsets TrueNorth d'IBM avec des neurones reliés entre eux par 130 millions de synapses sur 2 milliards de transistors[495]. Loihi est pour l'instant positionnée comme un composant destiné à de la recherche[496].



La puce Loihi, fabriquée en 14 nm, ne semble pas avoir évolué depuis 2017. Son packaging a cependant déjà traversé quatre générations : **Wolf Mountain** avec une carte comprenant 4 chipsets totalisant 512K neurones fin 2017, **Nahuru** avec 8 à 32 puces Loihi et jusqu'à 4 millions de neurones en 2018 (*ci-dessus à gauche*), **Kapoho Bay** sous forme de grosse clé USB avec un à deux Loihi fin 2018, pui **Pohoiki Beach**, comprenant 64 puces Loihi et 8 millions de neurones en juillet 2019, donc 8 milliards de synapses. Et puis enfin, en 2020, avec **Pohoiki Springs** qui comprend 768 chipsets Loihi et 100 de millions neurones répartis sur 24 cartes équipées d'un FPGA Arria, contenant chacune 32 chipsets, le tout consommant 500W[497]. On est encore loin des 100 milliards de synapses annoncées pour 2019 en juin 2018 et devant atteindre le niveau du cerveau d'un rat[498].

[495] Voir Introducing Loihi par Mike Davies, 2018 (30 slides) qui décrit bien l'architecture du chipset.

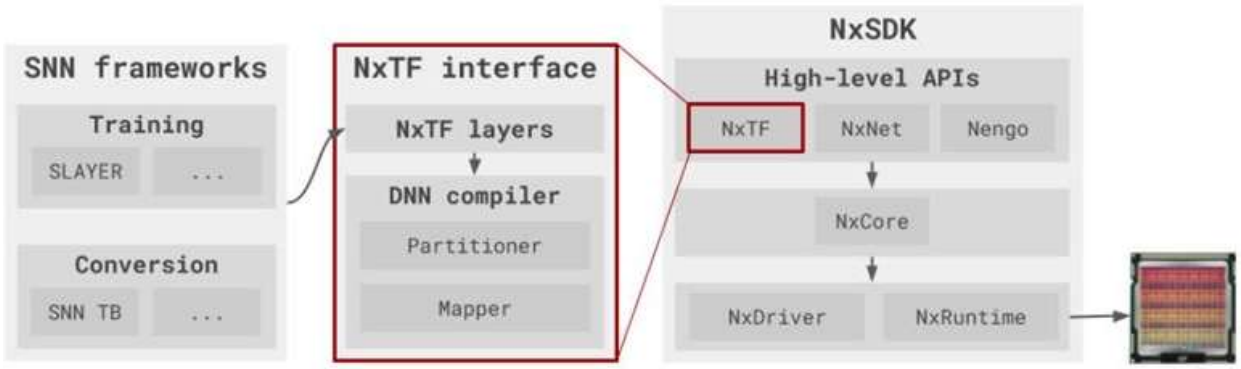
[496] Intel se distinguait sinon en intégrant un chipset Loihi dans un système de nez artificiel comprenant 72 capteurs. Voir Intel's Neuromorphic Nose Learns Scents in Just One Sniff par Samuel K. Moore, mars 2020.

[497] Voir Intel debuts Pohoiki Springs, a powerful neuromorphic research system for AI workloads par Kyle Wiggers, mars 2020.

[498] Relatées dans La puce Loihi d'Intel aussi performante qu'un cerveau de souris en 2019 de Mark Hachman en juin 2018. Loihi a été développé en partenariat avec l'institut Caltech.

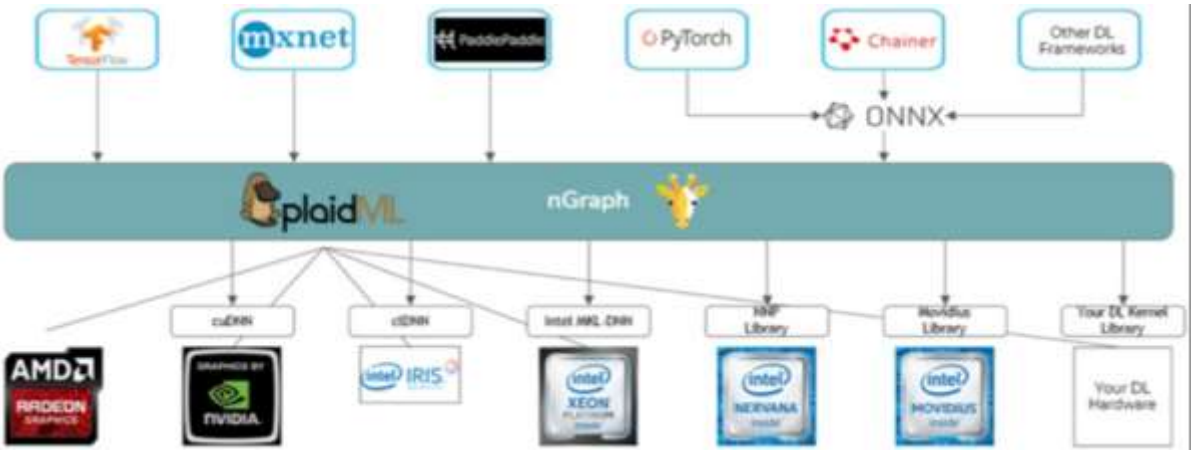
Le marketing d'Intel évoque des facultés d'apprentissage, en précisant qu'il supportera des réseaux de neurones récurrents, hiérarchiques et parcimonieux (sparse) et donc en particulier à tout ce qui correspond au traitement du langage et à l'analyse de flux de données temporels divers comme des électro-cardiogrammes[499].

Depuis janvier 2021, ces systèmes Loihi sont accompagnés d'un nouveau kit de développement et compilateur NxTF développé sur Keras facilitant le portage de réseaux de neurones à impulsion sur le chipset.



- La recherche d'Intel dévoilait sinon en 2020 un chipset d'inférence pour l'embarqué générant 617 TOPS par watt grâce au rapprochement de la mémoire et du calcul (Compute Near Memory - CNM) et à une réduction de la tension d'utilisation des transistors. Mais il ne s'agit pas encore d'un produit commercial. Les chipsets de tests étaient produits en 10 nm. 617 TOPS/watt est très impressionnant mais attention, ces TOPS sont appliqués à des BNN (Binary Neuron Networks), bref sur des neurones fonctionnant à 1 bit[500].

Pour alimenter toutes ces architectures matérielles disparates, Intel propose plusieurs outils de développement qui s'insèrent entre les frameworks leaders du marché comme TensorFlow et le matériel, y compris issu d'autres fournisseurs tels que Nvidia.



Intel nGraph est une bibliothèque C++ de compilation d'application de deep learning compatible avec le format ONNX qui permet d'exécuter des modèles de deep learning sur diverses architectures matérielles fixes et mobiles. nGraph supporte TensorFlow, PyTorch et MXNet.

[499] Tandis que les chipsets TrueNorth d'IBM ne gèrent pas l'apprentissage et ne font qu'exécuter les modèles neuronaux déjà entraînés, ici, le processeur est capable d'apprentissage et dans les modes supervisés, non supervisés et par renforcement.

[500] Voir Intel unveils array of chip research focused on edge data processing par Matt Hamblen, juin 2020.

Il est conçu pour optimiser les graphes de calculs pour exécuter l'entraînement et les inférences de réseaux de neurones[501].

plaidML est issu de l'acquisition en août 2018 de la startup **Vertex.IA** (USA), un petite équipe de 7 personnes[502], un framework open source de deep learning pour l'embarqué.

OpenVino est un SDK d'applications de reconnaissance d'images qui permet de les déployer dans les objets connectés et en particulier sur les chipsets Movidius[503] mais aussi sur les ordinateurs équipés des chipsets Core de la génération Ice Lake, gravés en 10 nm.

BigDL est une bibliothèque qui permet de distribuer des applications de deep learning sur des clusters de serveurs et d'exploiter Spark et Hadoop.

Intel faisait enfin l'acquisition de **SigOpt** (2014, USA, $8,7M) en octobre 2020. La société développé des outils exploitant du machine learning pour optimiser les paramètres d'un modèle de machine learning ou de simulation, en exploitant des optimisations bayésiennes. Cela fait penser à un modèle évolutionnaire. Il supporte notamment les frameworks TensorFlow, SciKit Learn, PyTorch, Keras et Mxnet[504].

Chipsets serveurs

De nombreux autres acteurs du marché s'activent pour créer des chipsets de deep learning destinés aux serveurs. Ce sont essentiellement des startups. On y trouve une grande variété d'architectures matérielles qui visent en général à optimiser le lien entre les calculs de tenseurs et la mémoire. Comment comparer ces différents processeurs ? Avec des benchmarks pardi ! D'où l'intérêt du lancement en mars 2020 du consortium **MLCommons** rassemblant des concepteurs de chipsets comme Graphcore et Kalray ainsi que des laboratoires de recherche et qui supervise les spécifications des benchmarks MLPerf.

Faisons maintenant le tour de quelques-uns des principaux chipsets de machine learning destinés aux serveurs :

- **Graphcore** (2016, UK, $722M) commercialise le Intelligence Processing Units (IPU), un chipset adapté à l'exécution d'applications de deep learning côté entraînement et inférence qui comprendrait 1000 cœurs. Ils ciblent notamment le marché autonomobile. En 2020, Graphcore annonçait la seconde génération de son IPU, le GC200, comprenant 59,4 milliards de transistors gravés en 7 nm, un peu plus que le Nvidia A100 (54 milliards) sur 823mm^2. Il lançait la M2000, un serveur 1U équipé de quatre de ces processeurs et 439 Go de mémoire partagée délivrant 1 PFLOPS ainsi que le rack IPU-POD$_{64}$ qui contient 16 serveurs M2000. Ces machines peuvent être installées en grappes pour consolider jusqu'à 64 000 IPU. Ces machines seront déployées à l'Université d'Oxford, au Lawrence Berkeley National Laboratory du DoE et chez Oxford Nanopore. Ces machines sont supportées par le SDK Poplar, lui-même supporté par les principaux frameworks de deep learning du marché comme TensorFlow et PyTorch ainsi que par le format pivot ONNX. Depuis janvier 2021, le rack IPU-POD de Graphcore est proposé sous la forme de services en cloud par **Cirrascale Cloud Services** (USA), devant monter en puissance jusqu'à des versions IPU-POD$_{64}$.

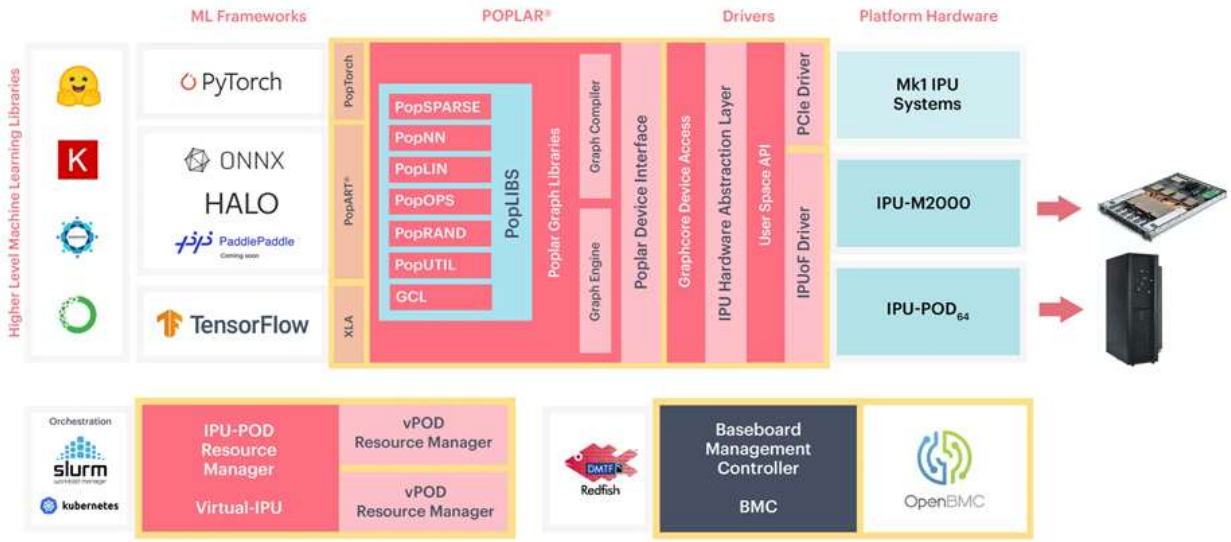
---

[501] Voir Unlocking Next-Generation Performance with Deep Learning Compilers par Adam Procter, Adam Straw et Robert Earhart, 2018 (45 slides) et Intel nGraph: An Intermediate Representation, Compiler, and Executor for Deep Learning, 2018 (3 pages).

[502] Il est sur GitHub ici : https://github.com/plaidml/plaidml. Le framework est particulièrement adapté aux ordinateurs tournant sous MacOS et Windows avec cartes graphiques AMD, Nvidia et les GPU intégrés dans les processeurs Intel. C'est en fait une surcouche qui supporte les applications développées avec le framework Keras ainsi qu'avec ONNX, Open Neural Network Exchange, un framework open source de bas niveau de réseaux de neurones, créé par Microsoft, Amazon et Facebook. Il supporte Caffe2, PyTorch, MXNet et le Microsoft CNTK. PLAID.ML est adapté aux réseaux de neurones convolutifs (CNNs) et à mémoire (LSTM).

[503] Voir la présentation Inference with Intel, 2018 (44 slides).

[504] Voir Intel acquires SigOpt, a specialist in modeling optimization, to boost its AI business par Ingrid Lunden, octobre 2020.



Ils publiaient en décembre 2020 un benchmark du M2000 montrant une équivalence de performance entre un cluster équipé de 8 chipsets M2000 ($259K) et 16 serveurs Nvidia DGX-A100 ($3M) pour des tâches d'entraînement. Les gains dans les inférences semblent encore plus importants aussi bien sur des modèles de reconnaissance d'images que de traitement du langage.

- **Groq** (2017, USA, $62,3M) est une startup lancée par des anciens de Google qui avaient participé à la conception de leurs TPU. Leur chipset pour serveur est censé générer 400 TFLOPS avec une puissance de calcul de 8 TFLOPS par Watt. Il cible aussi bien l'entraînement que les inférences de réseaux de neurones. Groq livre depuis 2020 des cartes d'accélération PCI et leur SDK associé. Ils évoquent la notion de « batch size 1 », qui correspond à l'usage de batches d'entraînement de réseau de neurones avec un seul objet de la base d'entraînement (au lieu de leur mutualisation avec de la descente de gradient réalisée en moyennant l'erreur de plusieurs objets de la base d'entraînement). Dans la lignée des premiers TPUs, ils utilisent aussi une grande matrice 320x320 qui n'est pas sans rappeler les matrices de 256x256 de la première génération de TPU. Le calcul matriciel est complété par des unités de traitement vectoriel avec des fonctions linéaires et non linéaires, utiles pour les fonctions d'activation des réseaux de neurones comme les sigmoïdes.

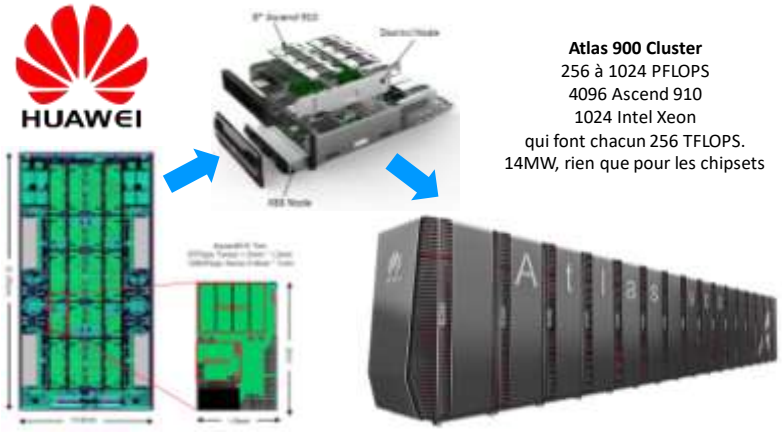
La mémoire cache SRAM est de 220 Mo avec une vitesse de transfert interne de 80 To/s. Ils optimisent la performance de l'accès à la mémoire avec cette SRAM et en évitant l'usage de mémoire cache à plusieurs niveaux comme chez Nvidia (avec leurs caches L1, L2 et L3)[505]. L'ensemble est packagé dans des serveurs 4U avec 8 cartes Groq TSP100 qui contiennent chacune un seul chipset Groq. Cela délivre 6 PETAOP/s en Int8. Le serveur comprend aussi un CPU AMD Epyc Rome. Le tout consomme 3,3 kW. On peut ranger 9 de ces systèmes par rack, ce qui fait 29,7 kW de consommation électrique par rack, soit l'équivalent de deux puces Cerebras.

[505] Voir Groq Shares Recipe for TSP Nodes, Systems par Nicole Hemsoth, septembre 2020.

- **Bitmain** (2013, Chine, $764M) est une licorne chinoise bien connue, à l'origine des chipsets de mining de Bitcoins, les AntMiner 9 qui dominent le marché. Elle a aussi développé un chipset d'IA, le Sophon BM1680 (*ci-contre*), qui sert à la fois à l'entraînement et à l'inférence de réseaux de neurones, voisin des TPU de Google. Il est destiné aux applications de traitement de l'image comme du langage, et notamment aux systèmes de vidéo surveillance dont la Chine rafole comme l'illustre le positionnement de nombreuses startups du pays (SenseTime, Face++, …).

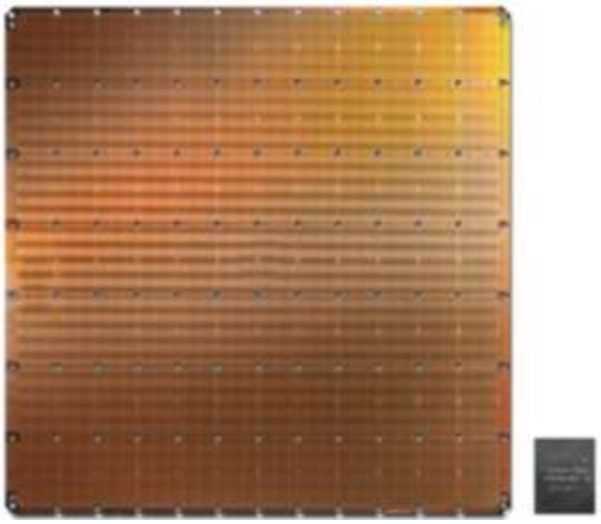
- **Tachyum** (2016, USA/Slovaquie, $17M) développe le Prodigy Universal Processor Chip, un processeur à 64 cœurs basse consommation pour les data centers.

- **Huawei** lançait en août 2019 son processeur pour entraînement d'IA sur serveur Ascend 910 supportant 256 TFLOPS et 512 TOPS sur des entiers 8 bits avec une consommation maximale de 310W[506]. Il gère des calculs vectoriels et de cubes de 16^3 (4096) nombres flottants. Ils sont intégrés dans le cluster Huawei Atlas 900 qui cumule jusqu'à 1024 PFLOPS[507].



- **Cerebras Systems** (2016, USA, $112M) est une étonnante startup qui propose un énorme ASIC dédié notamment à l'entraînement de réseaux de neurones qui optimise les calculs de matrices faiblement denses (avec beaucoup de zéros)[508].

  La startup a été créé par des anciens de SeaMicro, une startup constructeur de serveurs à basse consommation acquise par AMD en 2012 pour $357M, complétée récemment par un dirigeant d'Intel, Dhiraj Mallick. En août 2019, la startup révélait ce qu'elle concevait en douce : le plus grand chipset au monde, le Cerebras Wafer Scale Engine. C'est en fait une matrice de 7*12 chipsets gravés sur un wafer et reliés entre eux, totalisant 1,2 trillions de transistors, 18 Go de SRAM et 400 000 cœurs SLAC (Sparse Linear Algebra Cores) sur 8,5 pouces x 8,5 pouces. Les défauts de gravure statistiquement élevés sur une telle surface sont gérés en court-circuitant les cœurs défectueux, mais qui doivent être isolés.



[506] Le tout après avoir lance un programme sur l'IA pour les développeurs couvrant tout des smartphones aux serveurs. Voir Huawei Releases the AI Developer Enablement Program, octobre 2018.

[507] Voir DaVinci: A Scalable Architecture for Neural Network Computing, gère des matrices de 163 flottants par Heng Liao et al, 2019 (44 slides).

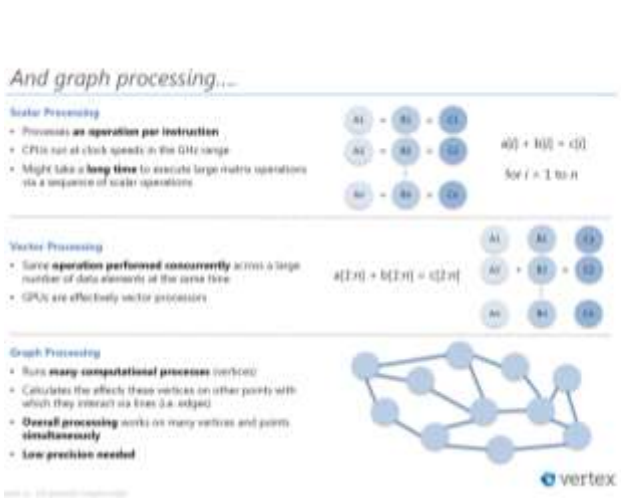
[508] Voir un peu plus de détails à leur sujet dans Le plus grand processeur d'IA, Olivier Ezratty, août 2019.

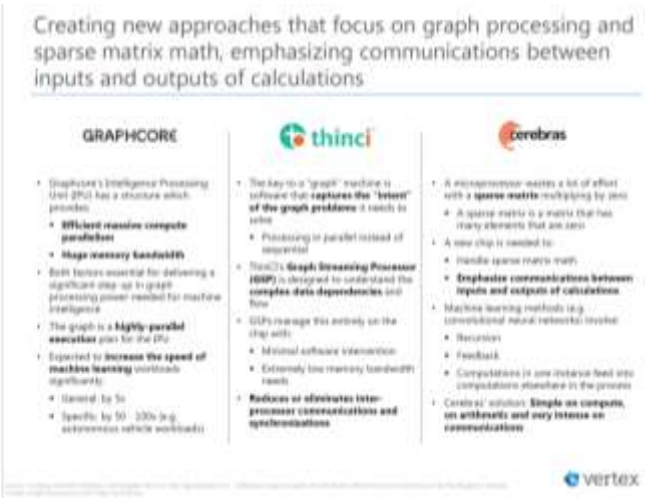
Le tout consomme 15 kW pour 4,8 TOPS par Watt et est refroidi par eau sur l'ensemble de sa surface, « perpendiculairement au composant » tout en tenant compte des phénomènes de dilatation dans les jonctions métal entre les chipsets du wafer. C'est une première et c'est une approche véritablement originale. On verra à l'usage !

Fin 2019, il était intégré dans un serveur CS-1 testé par le Département de l'Energie US[509]. Suivait en 2020 la création d'un serveur baptisé Neocortex exploitant deux CS-1 couplés à un système de mémoire et de stockage HPE, déployé par le Pittsburgh Supercomputing Center (conjoint entre les Universités Carnegie-Mellon et de Pittsburgh)[510]. Il est notamment mis à disposition de la recherche médicale.

Un benchmark publié en octobre 2020 montrait que le CS-1 était aussi très performant en calcul scientifique dans un cas de modélisation de dynamique des fluides[511]. Il était 200 fois plus rapide que le supercalculateur Joule utilisant 16 384 coeurs de CPU consommant 450 kW.

- **Amazon** a développé son propre chipset ASIC d'inférences pour serveurs, dénommé Inferentia, fin 2019. Ce chipset est construit autour de quatre cœurs couplés à un seul niveau mémoire cache puis à de la mémoire DRAM. Ces cœurs permettent de réaliser du calcul matriciel (« systolic arrays »). Mais Amazon ne fournit pas plus de détails techniques sur l'architecture interne de ces chipsets. Ils supportent TensorFlow, PyTorch, MXNet et le format pivot ONNX. Il est proposé dans l'offre en cloud EC2 et est supporté par leur environnement de développement SageMaker. Le chipset est conçu pour offrir un très faible temps de latence. Les puces Inferentia délivrent 120 TOPS en int8 (entiers 8 bits). Elles sont assemblées par paquets de 16 dans des instances inf1, et couplées à des CPU Intel Xeon. Ces puces ont été conçues par Annapurna Labs, une startup israélienne acquise par Amazon en 2015.
- **Blaize** (2010, USA, $65M), anciennement ThinCI, développe ses Graph Streaming Processor ASIC pour de la vision artificielle. Le chipset optimise l'organisation des graphes à calculer.





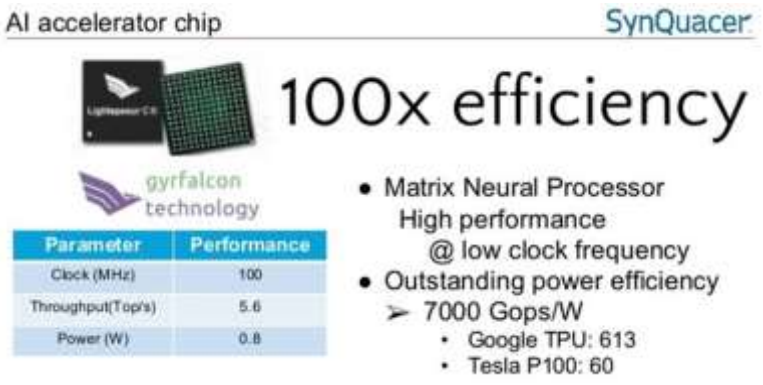
[509] Voir Le CS-1, super-ordinateur dédié à l'IA, est prêt à révolutionner la recherche par Thomas Burgel, novembre 2019 et Argonne National Laboratory Deploys Cerebras CS-1, the World's Fastest Artificial Intelligence Computer | Argonne National Laboratory, novembre 2019.

[510] Voir Cerebras' Wafer Scale Engine Scores a Sale: $5m Buys Two for the Pittsburgh Supercomputing Center par Dr. Ian Cutress, juin 2020 et Neocortex Will Be First-of-Its-Kind 800,000-Core AI Supercomputer par By Tiffany Trader, juin 2020.

[511] Documenté dans Fast Stencil-Code Computation on a Wafer-Scale Processor par Kamil Rocki et al, octobre 2020 (12 pages). Le papier décrit au passage la structure de calcul des unités de traitement du CS-1. Chacun des cœurs du CS-1 réalise des opérations simples de type ax+y, et accède à 48 Ko de mémoire SRAM.

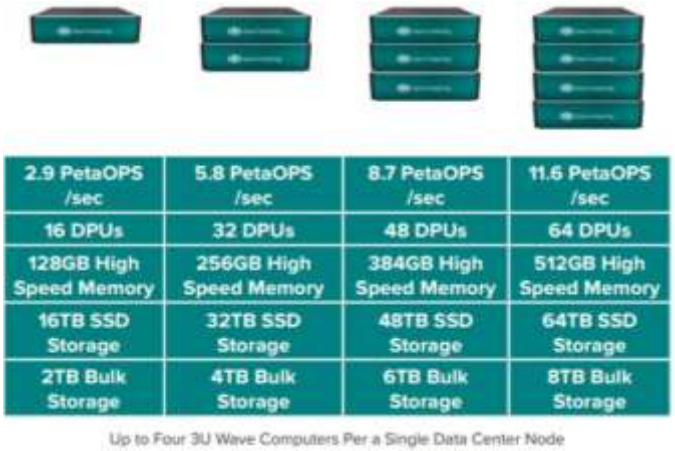
- **Gyrfalcon Technology Inc** ou GTI (2017, USA) est sorti du bois en septembre 2017 avec deux chipsets d'inférences ASIC à basse consommation, l'un pour les serveurs et l'autre pour les objets connectés. La version serveur (Lightspeeur 280x AI Accelerator) est intégrée dans des cartes à 16 composants[512]. Le Lightspeeur 5801 de la quatrième génération est un ASIC produit en technologie 28 nm chez TSMC. C'est le premier accélérateur d'IA du genre qui permet une sérieuse accélération des traitements. Son efficacité énergétique est très bonne avec 12,6 TOPS/watt.

  Il est équipé de gestionnaire de matrices 168x168 avec plus de 28 000 cœurs. Ce genre de chipsets peut équiper des smartphones pour l'ajout de fonctionnalités diverses : création de bokeh d'arrière plan, augmentation de la résolution, segmentation d'images, détection d'objets ou amélioration d'images en basse lumière.



- **SambaNova Systems** (USA, $206M) développe aussi un chipset d'inférences pour serveurs qui cherche à dépasser la performance des V100 de Nvidia et cible le traitement de données. La startup a aussi obtenu $8M de financement de la DARPA.
- **ThinkForce Electronic Technology** (2017, Chine, $68M) développe des chipsets de deep learning pour serveurs, basés sur des architectures multicœurs.
- **Baidu** présentait en juillet 2018 son chipset serveur **Kunlun** réalisé en 14 nm chez Samsung qui semble être généraliste, étant adapté aussi bien à la reconnaissance d'images qu'au traitement du langage. C'est visiblement dans un premier temps un FPGA. Il a une bande passante mémoire convenable de 512 Go/s et délivre 260 TOPS pour une consommation de 100W.
- **Enflame Technologies** (2018, Chine, $474M) développe des processeurs d'entraînement d'IA. Le Yunsui T11 génère 88 TFLOPS en BF16 avec 300W. Il comprend 16 Go de mémoire HBM2.
- **Wave Computing** (2010, USA, $203M) développait ses Dataflow Processing Units avec 16 000 cœurs produits en ASIC chez TSMC en 16 nm, dédiés à l'entraînement de réseaux de neurones.

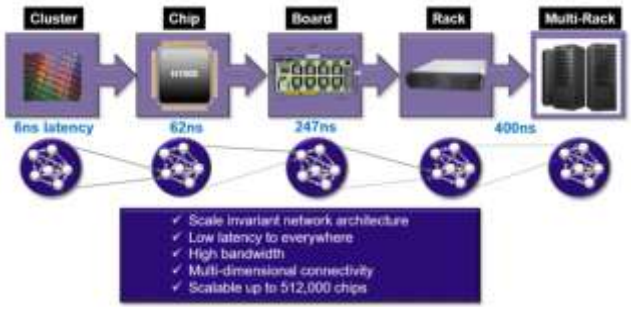
  Ces DPU étaient assemblés dans des serveurs par paquets de 16, donnant 128 000 cœurs. Ils n'utilisaient par contre que de la DRAM, bien moins performante que la mémoire HBM des GPU Nvidia et autres chipsets plus spécialisés[513].

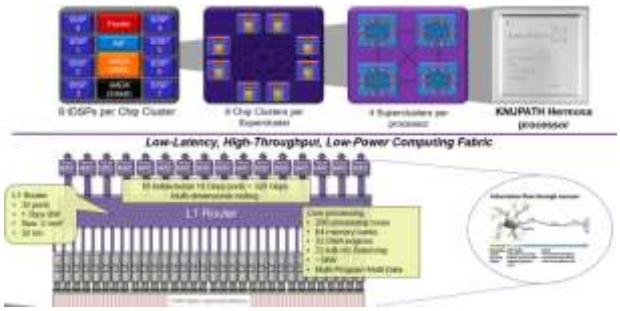


[512] Voir AI Accelerator Gyrfalcon Soars Post Stealth de Kevin Fogarty, novembre 2018.

[513] Source : AI processing requirements reveal weaknesses in current methods de Tom Simon, juillet 2018. Wave Computing avait fait l'acquisition de l'activité MIPS de l'Anglais Imagination Technologies. Elle est en redressement judiciaire (Chapter 11) depuis avril 2020.

- **Prefered Network** (2014, Japon, $147,2M) est une société de R&D de solutions de deep learning ciblant plusieurs marchés (automobile, industrie, santé, robotique, etc). Elle opère ses propres supercalculateurs, avec trois installations totalisant 2560 GPU et 200 PFLOPS. Ils développent leur propre accélérateur maison devant être déployé en 2020. Ils sont aussi à l'origine du framework de deep learning open source Chainer adapté au déploiement sur des grappes de clusters, exploitable sous Python et supportant notamment les GPU de Nvidia. Ils proposent aussi Optuna, une brique logicielle open source qui permet d'optimiser les hyperparamètres d'un réseau de neurones.
- **Pezy Computing** (2010, Japon) développe le chipset SC2 comprenant 2048 cœurs tournant à 1 GHz et générant 4 FLOPS per cycle par cœur et un total de 8 PFLOPS (en entier single-precision). Il est fabriqué en 16 nm chez TSMC.
- **KnuEdge** (2007, $100M) planche sur un chipset Knupath qui est basé sur la technologie LambaFabric qui permet l'alignement en parallèle de 512 000 unités assemblées dans des chipsets de 256 cœurs. L'offre comprend les chipsets KnuVerse dédié à la reconnaissance de la parole pour l'authentification, ainsi que les services en cloud Knurld.io permettant d'intégrer l'authentification vocale dans une application.





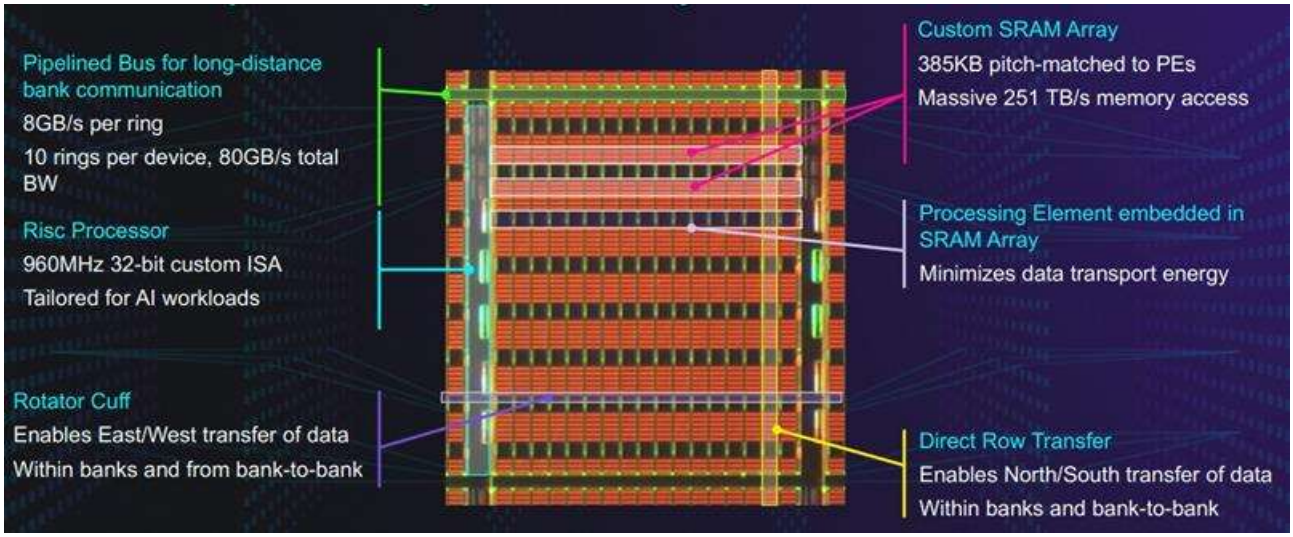
- **Alibaba** et **Tencent** ont aussi leurs chipsets serveurs de machine learning pour serveurs, aussi réalisés en FPGA[514]. Fin septembre 2019, Alibaba annonçait son nouveau chipset d'inférence pour serveur, Hanguang 800[515]. Il ne tournera que dans les serveurs cloud d'Alibaba. Le chipset a été conçu par leur filiale **T-Head**. Ils lançaient aussi le XuanTie 910 qui est dédié aux objets connectés, et basé sur un noyau RISC-V.
- **Qualcomm** annonçait au printemps 2019 une carte d'accélération pour le deep learning destinée aux serveurs dans le cloud, en format PCI, le Qualcomm Cloud AI 100. Mais avec une communication minimaliste : aucune spécification technique n'est véritablement disponible pour comprendre ce que ces cartes contiennent exactement. On sait juste que les cartes dédiées aux inférences de réseaux de neurones contiennent des composants ASIC gravés en 7 nm et qu'elles ont une capacité de calcul de 350 TOPS. Mais rien sur la mémoire qui est un point clé, sur le nombre se tenseurs, ni sur le type de fonction de calcul (tenseurs, vecteurs). Ce n'est pas bien sérieux. Mais Qualcomm n'est pas encore un acteur sérieux dans les datacenters et cela ne risque donc pas de changer. Ces cartes étaient annoncées pour 2020.
- **PQ Labs Inc** (USA) lançait en janvier 2020 sa « QuantaFlow AI architecture » au CES 2020. Il s'agit d'une architecture exploitant un processeur RISC-V, un "QuantaFlow Generator" et un « QF Evolution Space ». Elle est présentée comme capable de paralléliser les traitements un peu comme dans les transformations unitaires du calcul quantique. QuantaFlow simule une transformation de registres « qf-bits ».

[514] Voir Alibaba Launches Chip Company "Ping-Tou-Ge"; Pledges Quantum Chip, septembre 2018.

[515] Voir Announcing Hanguang 800: Alibaba's First AI-Inference Chip, octobre 2019 et Processing Near or In Memory For Deep Learning par Lide Duan, Alibaba, mai 2019 (37 slides).

Le « QuantaFlow Generator » convertit des données en entrée d'un faible espace dimensionnel dans un espace plus dimensionnel plus grand sur lequel il réalise des évolutions et transformations et avec un très fort parallélisme. Le système serait capable d'exécuter toutes sortes de réseaux de neurones plus efficacement que les GPGPU et ASIC dédiés au deep learning. Le gain de performance affiché serait de 10x sur des réseaux convolutifs type ResNet-50 par rapport à GPU Nvidia V100. La société est spécialisée dans les écrans multitouch. Je n'ai pas trouvé de littérature scientifique sur leur procédé, c'est donc sujet à caution.

- **Untether AI** (2017, Canada, $40M) propose un chipset optimisé pour l'inférence, le runAI200, qui est intégré par groupe de quatre dans des cartes TsunAImi PCIe. Il optimise le calcul en le rapprochant de la mémoire, grâce à 200 Mo de SRAM répartie dans 511 blocs mémoire. Le chipset comprend 18,4 milliards de transistors qui sont gravés en 16 nm. Il tourne à 720 MHz en mode normal et à 960 MHz en mode turbo. Le chipset génère 502 TOPS en int8 pour 8 TOPS/watts, ce qui est une bonne performance[516].



- **SiPearl** (France) est une startup crée par Philippe Notton dans le cadre du projet européen EPI (European Processor Initiative) et EuroHPC. Elle conçoit Rhea, un processeur destiné entre autre chose aux futurs supercalculateurs européens cofinancés par l'Union Européenne. Ce processeur utilisera 72 noyaux arm Neoverse Zeus[517], et des blocs de propriété intellectuelle d'origines variées comme Kalray.

  Le champ principal d'application sera le calcul scientifique, même si des applications de machine learning ne sont pas exclues, malgré l'absence de tenseurs dans le chipset. Il devrait être produit en 7 nm chez TSMC. La société doit trouver $100M de financement privé, qui doit être abondé à hauteur de $100M par l'Union Européenne[518].

Chipsets dans l'embarqué

Nous allons maintenant passer aux processeurs adaptés à l'exécution d'applications de deep learning dans l'embarqué, à la fois dans les smartphones et dans les objets connectés en général. L'offre y est bien plus abondante que sur serveurs.

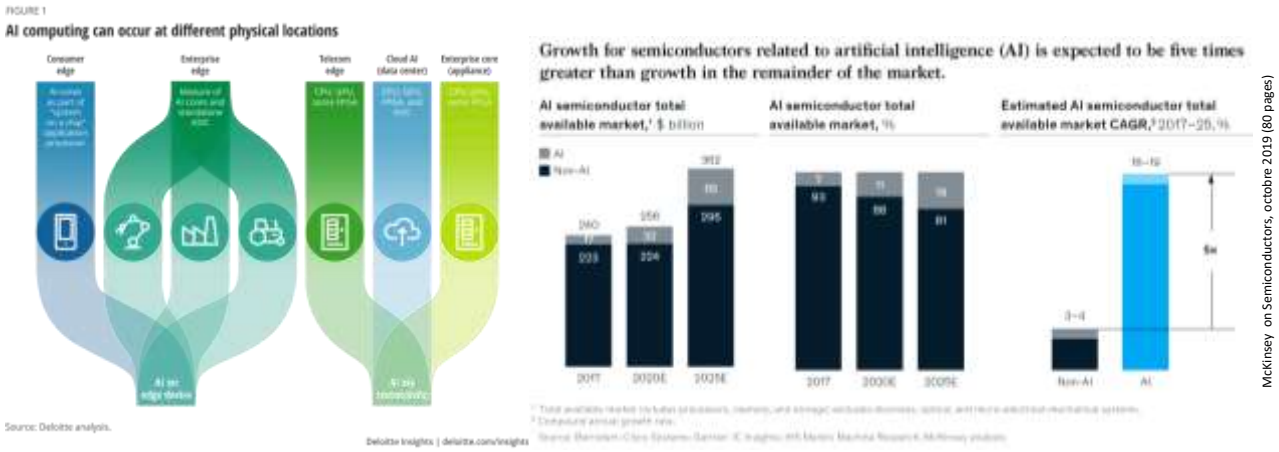
---

[516] Voir Server inference chips startup Untethered from AI data movement par Nicole Hemsoth, octobre 2020.

[517] Voir Arm Announces Neoverse V1 & N2 Infrastructure CPUs: +50% IPC, SVE Server Cores par Andrei Fumusanu, septembre 2020.

[518] Voir Accelerator - European Processor Initiative et SiPearl Lets Rhea Design Leak: 72x Zeus Cores, 4x HBM2E, 4-6 DDR5 par Andrei Frumusanu, 2020.

Une étude datant de fin 2019 de **Deloitte** présente bien la structuration de ce marché[519]. Une autre étude, de **McKinsey**, illustre le fait que le marché des semiconducteurs intègre progressivement des briques d'IA[520]. La part des composants d'IA passerait de 11% du marché en 2020 à 19% du marché en 2025. C'est une évolution naturelle qui fait écho à ce qui est en train de se passer dans le marché du logiciel où la majorité des acteurs y intègre maintenant des briques d'IA, surtout de machine learning. Dans les semi-conducteurs, il en va de même avec un mix - plutôt minoritaire - de composants 100% dédiés à l'IA et de composants hybrides intégrant des fonctions d'IA, comme les chipsets de smartphones, ou le récent chipset M1 qui équipe les Macintosh annoncés en novembre 2020.



Sur smartphone, le premier des chipsets d'IA en date est le Kirin 970 de **HiSilicon**, la filiale de semiconducteurs du chinois Huawei. Présentée à l'IFA 2017, il s'agissait d'un chipset mobile gravé en 10 nm par TSMC et comprenant 5,5 milliards de transistors.

Kirin 9000
NPU 5 nm
4 TOPS
Huawei Mate 40

Apple A14 Neural Engine
5 nm
11 TOPS
iPad Air 2020, iPhone 12

Qualcomm Snapdragon 865
Neural Engine 7 nm
15 TOPS
Samsung Galaxy S10, LG V50 Thinq et le G8 Thinq, Xiaomi Mi 9 et Mi3, Sony Xperia 1 (Snapdrahon 855).

Il comprenait un NPU (Neural Processing Units) fait de multiplicateurs de matrices 3x3 dédiés au traitement d'applications de deep learning comme la reconnaissance de la parole ou d'images. On l'a retrouvé dans de nombreux smartphones lancés depuis comme les Huawei Mate 10, Pmate 20 et Honor 10. Le NPU peut traiter 1,92 TOPS. Dans le Kirin 980, la puissance était doublée à 4 TOPS. Le NPU supporte Tensorflow, Tensorflow Lite et Caffe/Caffe2. Ce chipset a été suivi en septembre 2018 du Kirin 980 qui double la puissance côté NPU, et est gravé en technologie 7 nm.

Avec 10,3 milliards de transistors, le Kirin 990 5G lancé en septembre 2019 comprend un NPU DaVinci comprenant trois cœurs, qui supportent des calculs entiers (INT8) et flottants (FP16) avec deux cœurs Ascend Lite (pour traitements complexes) et un cœur basse consommation Ascend Tiny (pour traitements légers, reconnaissance de visages et de la parole)[521].

[519] Voir Bringing AI to the device: Edge AI chips come into their own par Duncan Stewart et al, décembre 2019.

[520] Voir McKinsey on Semiconductors, octobre 2019 (80 pages).

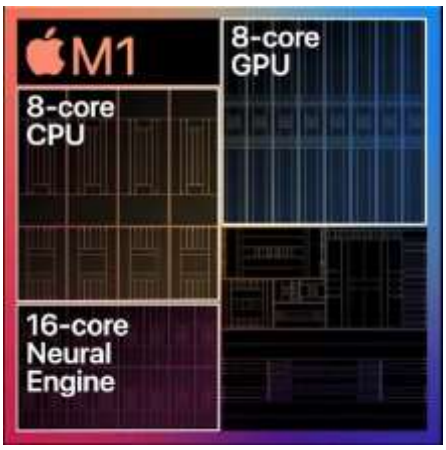
[521] Voir Huawei Introduces World's First Flagship 5G SoC in Challenge to Samsung and Qualcomm, septembre 2019.

Les Kirin 970 et 980 utilisaient une conception de circuit provenant de **Cambricon Technology** (2016, Chine, $200M[522]). A partir du Kirin 990, Huawei a repris son autonomie pour ce qui est de la conception de NPU.

En 2020, Huawei passait au Kirin 9000 produit chez TSMC en technologie 5 nm. Il intègre directement un modem 5G. Il comprend 15,3 milliards de transistors, à comparer aux 11,8 milliards de transistor de l'A14 d'Apple qui est aussi fabriqué en 5 nm et qui équipe les iPhone 12 lancés en octobre 2020. La taille du Kirin 9000 est aussi due au GPU Mali-G78MP24 et à ses 28 cœurs. Le Kirin 9000 comprend un nouveau NPU de puissance non précisée. Ce chipset équipe le smartphone Mate 40 lancé en octobre 2020. Mais sa fabrication a dû être stoppée en septembre 2020 à cause des sanctions américaines contre Huawei qui s'appliquent à TSMC qui utilise des machines de production d'origine américaine, comme celles d'Applied Materials.

Une semaine après l'annonce du Kirin 970 par Huawei en 2017, **Apple** lançait sa salve d'iPhones annuelle, les 8 et X. Ceux-ci intégraient aussi pour la première fois une fonction neuromorphique sous la forme d'un coprocesseur dénommé A11 Bionic Neural Engine. Il tourne à 900 MHz. On sait sans surprise qu'il est exploité par SIRI et par les fonctions de reconnaissance d'images comme le login exploitant une vue 3D du visage. Le chipset A12 équipant les iPhone XR lancés en septembre 2018 doublait la puissance du NPU à environ 3-4 TOPS. Le chipset A13 de septembre 2019, gravé en 7 nm, qui équipe les iPhone 11 est encore plus puissant et atteindrait 5 TOPS. Son NPU représentait 4,7% du SoC et environ 400 millions de transistors. Le A14 gravé en 5 nm lancé en 2020 galoperait à 11 TOPS pour sa partie NPU qui comprend 16 cœurs de nature non précisée, occupe environ 6,8% de la surface du chipset et comprendrait environ 812 millions de transistors. Il équipe les iPad lancés en septembre 2020 et les iPhone 12 lancés en octobre 2020.

Dans cette lignée, Apple lançait en novembre 2020 trois nouveaux Macintosh portables et de bureau (MacBook Air, MacBook Pro et Mac Mini) intégrant une puce M1 à base de noyaux arm, remplaçant les Intel Core qui les équipaient depuis 2005. Le M1 est fabriqué en 5 nm comme le A14 et comprend 16 milliards de transistors, un peu plus que les 11,8 milliards de transistors du A14. Sa mémoire limitée à 16 Go est partagée par tous les modules intégrés, CPU, GPU et NPU, ce qui est censé accélérer certaines fonctions comme pour jouer des vidéos. Il comprend 8 cœurs arm en architecture big.LITTLE (quatre rapides et quatre lents), un GPU à 8 cœurs et 25 000 threads et un NPU à 16 cœurs.



Le tout doit permettre d'obtenir une autonomie de 15 à 18 heures sur un laptop neuf[523] ! Apple annonçait des gains de performance allant jusqu'à un ordre de grandeur par rapport aux classiques Intel Core i7. On pourra juger sur pièces lorsque ces Macintosh seront testés de manière indépendante et avec des logiciels adaptés au M1 au lieu de fonctionner en mode d'émulation Intel.

Qualcomm a été le dernier des grands concepteurs de chipsets de smartphones à se mettre aux NPU. Il a fallu pour cela attendre la sortie des Snapdragon 855 gravés en 7 nm en décembre 2018, suivis du 855+ qui améliorait la partie CPU et GPU. Ils contiennent le Tensor Xccelerator, un NPU de nature non précisée atteignant 7 TFLOPS. Il fonctionnerait en nombres entiers 16 bits vs 8 bits pour ses concurrents. Il gère des opérations matricielles et non linéaires pour les fonctions d'activation de neurones.

---

[522] Ces $200 ont été levés auprès de SDIC, un investisseur public chinois qui ressemble à notre Bpifrance. Fin 2017, Cambricon annonçait lancer d'ici 2019 la fabrication de son propre chipset en 16 nm chez TSMC, les MLU100 et MLU200 (Machine Learning Units) dédiés aux serveurs.

[523] Voir Apple launches three "mind-blowing" Macs with its M1 chip and Big Sur OS par Matt Hamblen, novembre 2020.

Ce NPU est intégré dans le DSP Hexagon 690 du chipset (685 pour le 855+). Le Snapdragon 855 comprend aussi le Spectra 380 ISP qui contient la Computer Vision Engine (CV-ISP) qui est dédiée au traitement de vidéos en 60p. Cela lui permet de détecter et détecter, classifier et segmenter les objets dans les images en basse consommation. Cela supporte par exemple les fonctions de floutage automatique de l'arrière-plan.

Avec le 855, Qualcomm gérait les applications d'IA essentiellement via le Snapdragon Neural Processing Engine SDK supportant d'un côté les principaux frameworks de deep learning du marché (Tensorflow, Caffe, Caffe2, ONNX et les API Android Neural Networks) et de l'autre, exploitant les différentes composantes des chipsets Snapdragon, les cœurs Kryo, le GPU maison Adreno et les DSP Hexagon et leurs unités de traitement vectorielles voisines de ce que l'on trouve dans les processeurs serveur Intel Xeon Phi. La bibliothèque Hexagon Neural Network permet d'exécuter des logiciels de deep learning directement sur les processeurs vectoriels Hexagon, notamment pour des réseaux convolutifs.

Qualcomm s'appuie notamment sur les compétences des équipes de **Scyfer** (2013, Pays-Bas), une startup issue de l'Université d'Amsterdam acquise en 2017 et spécialisée dans les développements logiciels en machine learning. Ils font aussi appel à **Brain Corp** (2009, USA, $125M) dans lequel Qualcomm Ventures a investi et pour de la R&D externalisée dans la vision artificielle. Conscients de leur besoin de se créer un écosystème de partenaires, ils lançaient en décembre 2018 un fonds d'investissement dans les startups de l'IA avec une première prise, **AnyVision** (2015, Israël, $74M), avec leur solution Better Tomorrow dédiée à la reconnaissance de visages, corps et objets dans la foule. Une solution plutôt destinée aux services de sécurité.

En décembre 2020, Qualcomm lançait le Snapdragon 888 succédant au 765G. Il intègre un modem 5G (X60) supportant les ondes millimétriques et tout un tas d'autres fonctionnalités étendues, notamment côté traitement de l'image. Pour ce qui est du machine learning, ses tenseurs délivrent 26 TOPS ce qui semble être le record en date. Ils sont intégrés dans le DSP Hexagon 780 qui unifie autour d'une même mémoire les modules de calcul scalaire (sur des variables discrètes), vectoriel (matrices à une dimension) et tensoriels (matrices à deux dimensions ou plus).

Au CES 2018, le Chinois **Rockchip** sortait son premier processeur embarqué RK3399Pro intégrant un NPU, atteignant 2,4 TOPS, au niveau du Kirin 970 de HiSilicon.

Le Taïwanais **MediaTek** faisait de même en annonçant sa plateforme NeuroPilot, une "AI processing unit" (APU) associée à un SDK NeuroPilot qui supporte les habituels outils de l'IA tels que TensorFlow, Caffe et Amazon MXNet. Les marchés visés sont les smartphones et l'automobile. Comme c'est souvent le cas, l'annonce ne précisait pas les fonctions mathématiques mises en œuvre dans leur APU, ce qui est bien dommage mais risque d'être courant.

Dans l'embarqué, nous avons aussi une belle brochette d'offres disponibles ou en cours de développement :

- **Mythic** (2012, USA, $85,2M) conçoit des chipsets pour micro-ordinateurs qui sont optimisés du côté de la gestion de la mémoire et avec une interface PCIe. Ils utiliseraient une méthode associant numérique et analogique pour accélérer les inférences dans les réseaux de neurones. Le chipset sera notamment supporté par TensorFlow[524].
- **NeuroBlade** (2017, Israël, $27,5M) propose une puce d'IA pour l'embarqué et pour serveurs qui optimise les traitements de plusieurs réseaux de neurones simultanément et optimise l'accès à la mémoire. À peine sortie du bois, elle ne communique pas beaucoup sur sa technologie.

[524] Voir Mythic nets $40M to create a new breed of efficient AI-focused hardware de Matthew Lynley, mars 2018.

- **Horizon Robotics** (2015, Chine, $700M) est spécialisée dans la vision artificielle pour la conduite autonome, la vidéosurveillance et les applications dans le retail pour y analyser et optimiser le parcours des clients (vidéo). Leur offre s'appuie sur divers composants matériels et logiciels. Leur chipset Sunrise sert à la reconnaissance de visages dans des vidéos, supportant une entrée en Full HD et le suivi simultané de 200 objets avec une consommation de seulement 1,5 W. Cela permet de simplifier la conception de caméras de surveillance interprétant elles-mêmes les images captées. Un autre chipset, le Journey suit les trajectoires de 20 classes d'objets comme les piétons, les vélos, les voitures et la signalisation routière (vidéo). Ils fournissent aussi des chipsets pour les haut-parleurs connectés de Xiaomi. Ils utilisent un jargon pas facile à décrypter avec leur Brain Processing Unit (BPU) qui décrit l'architecture des chipsets puis l'Elastic Tensor Core et le MIMD (Multiple Instruction Multiple Data) qui correspondent à leur méthode de parallélisation des traitements dans leur chipset. Les chipsets sont fabriqués avec un mix de FPGA et d'ASIC qui n'est pas bien clair. Les composants sont fabriqués chez l'incontournable TSMC. Leur plateforme logicielle Hugo comprend trois générations d'architecture : Gauss, Bernoulli et Bayes selon le niveau de sophistication des réseaux de neurones utilisés, Matrix étant bâti sur le niveau Bernoulli.

Pour les véhicules autonomes de niveaux 3 et 4, Horizon Robotics propose le système Matrix équipé de leurs chipsets et qui analyse les images des caméras et LIDARs du véhicule. Consommant seulement 31 W, il est refroidit passivement (*ci-dessus à droite*). Mais il faut trois Matrix pour analyser l'environnement à 360°, exploitant 12 caméras (4 grand angle et 8 classiques). La startup développe aussi NavNet qui gère la localisation dynamique des véhicules en s'appuyant sur des cartes 3D partagées dans le cloud. Cela rappelle un projet équivalent de Mobileye.

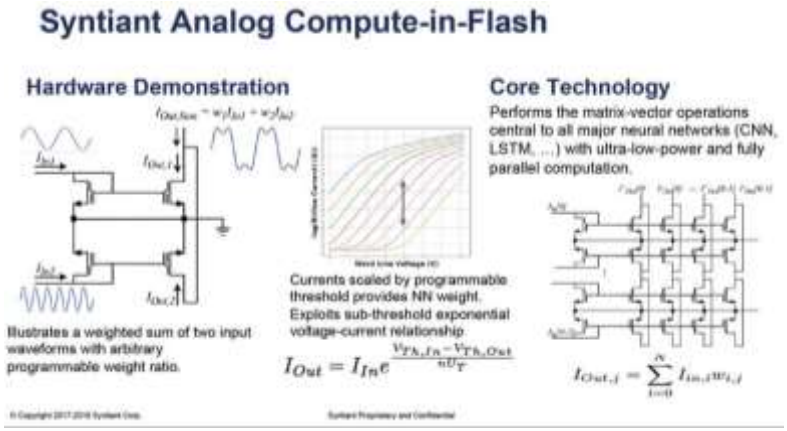
- **Vathys** (2015, USA, $120K) développe un chipset censé être 10 fois plus rapide que les concurrents en optimisant le mouvement des données. Ils en sont pour l'instant à l'état du concept[525].

- **Hailo Technologies** (2017, Israël, $87,9M) développe le Hailo-8, un processeur neuronal censé apporter la performance des data centers dans des applications embarquées, à basse consommation et à refroidissement passif, avec 26 TOPS[526]. Comme c'est souvent le cas, cela passe par une architecture qui rapproche la mémoire des unités de traitement, dénommée NMP (Near Memory Processing). Les marchés visés sont variés : véhicules autonomes, vidéosurveillance, drones, maintenance industrielle, etc.

[525] Voir les détails dans Vathys Petascale Deep Learning on a (Single) Chip, 2017 (28 slides).

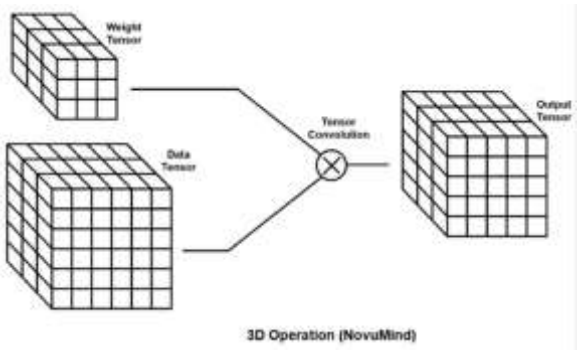
[526] Voir Hailo challenges Intel and Google with its new AI modules for edge devices par Frederic Lardinois, septembre 2020 et Edge-AI Startup Gets $60-million, Preps for Mass Production par Samuel K. Moore, mars 2020.

- **Syntiant** (2017, USA, $5,1M) développe le chipset Neural Decision Processors (NDP) qui intègre une sorte de mémoire Flash analogique à côté de ses fonctions d'inférence de deep learning organisées sous forme de multiplicateur de matrices par des vecteurs, le tout fonctionnant à basse consommation grâce à des synapses avec une basse précision de 3 à 5 bits.



  Le chipset génère 20 TOPS/watt. Il vise bien entendu les applications mobiles en tant que co-processeur[527].

- **NovuMind** (2015, USA, $15,2M) développe ses chipsets NovuTensor, dédiés aux inférences de réseaux de neurones convolutifs pour la reconnaissance d'images. Ils génèrent 15 TFLOPS pour 5 Watts et 3 TFLOPS par Watt, ce qui semble être maintenant la norme. Le NovuTensor est disponible sous la forme de composant ou dans une petite carte PCIe qui peut s'installer dans un serveur. L'ensemble était en bêta en date de septembre 2018.



  Sa sauce magique ? La gestion de tenseurs (matrices) à trois dimensions. Le tout est complété par les outils de développement de NovuBrain et une offre d'entraînement de réseau de neurones en cloud s'appuyant sur des V100 de Nvidia.

- **CogniMem** (2011, France/USA) propose le CM1K est un chipset ASIC intégrant un réseau de 1024 neurones stockant chacun 256 octets qui sert aux applications de reconnaissance des formes. Ne coûtant que $94, il est notamment utilisé dans la **BrainCard**, issue de la start-up franco-américaine, **General Vision** qui commercialise des « blocs d'IP » pour créer des processeurs neuromorphiques, avec ses NeuroMem. Cette technologie est aussi intégrée dans les processeurs d'objets connectés Curie d'Intel (avec 128 neurones, mais abandonnés par ce dernier en juillet 2017). L'ensemble sert principalement aux applications de vision artificielle dans les systèmes embarqués.

- **DeePhi Tech** (2016, Chine, $40M) prévoyait de sortir des chipsets neuromorphiques en 2018, l'un pour le cloud et l'autre pour l'embarqué et spécialisés dans la reconnaissance d'images. Leur DPU (Deep Learning Processor Unit) est un FPGA complété par le SDK Deep Neural Network Development Kit (DNNDK). Cette spin-off issue d'un partenariat entre l'Université de Tsinghua et celle de Stanford vise le marché des caméras de surveillance et des robots. Leur architecture de composant optimise le temps d'accès aux paramètres des réseaux de neurones en mémoire vive[528]. DeePhi a été acquis par l'Américain Xilinx en juillet 2018 qui propose par ailleurs son DNN Processor pour les FPGA Xilinx avec ses outils logiciels associés[529]. Xilink est le numéro un mondial des FPGA, devant Intel.

---

[527] Le schéma provient de la présentation Analog Computers for Deep Machine Learning de Jeremy Holleman, le CTO de Syntiant (19 slides).

[528] Voir Accelerating Inference at the Edge par Song Han, presentation à la conférence HotChips 2018 (102 slides).

[529] Voir Xilinx DNN Processor, 2018 (20 slides).

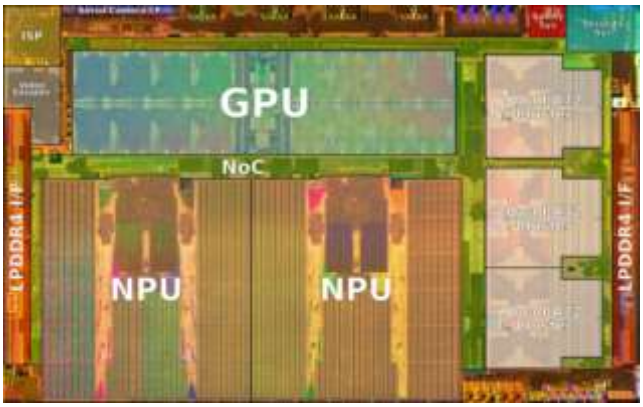
- **Eta Compute** (2015, USA, $8M) lançait en octobre 2018 son chipset d'IA Tensai à très basse consommation censé permettre l'autoapprentissage. Il sert aussi bien à la classification d'images qu'au traitement du langage et à la commande vocale. Il supporte les architectures de réseaux convolutifs aussi bien qu'à neurones à impulsions (spiking neurons). L'autoapprentissage consiste à détecter des anomalies dans les données fournies. Il supporte les frameworks TensorFlow et Caffe.
- **AnotherBrain** (2017, France, 30M€) est une startup lancée par Bruno Maisonnier, le fondateur d'Aldebaran Robotics. L'architecture de son chipset n'est pas documentée à ce stade. Le concept serait très différent de tous les autres chipsets neuromorphiques, avec, comme avantage, un processus d'entraînement plus rapide et nécessitant moins de données et d'énergie. La startup vise notamment des applications industrielles et dans l'automobile. Le principe a l'air d'être de l'apprentissage à base de « handcrafted features » pour détecter des anomalies dans des images[530].
- **Renesas** propose sa **s**érie de micro-processeurs basse consommation RZ/V qui intègrent un accélérateur de machine learning pour la vision artificielle, le DRP-AI (Dynamically Reconfigurable Processor). Consommant 4W, il est refroidi passivement. Une version haut de gamme permet l'analyse d'images 4K/30p en temps réel. Il est notamment adapté à la lecture de codes-barres, à la reconnaissance d'iris et de visages. Son rendement énergétique de 1 TOPS/W est classique. Il comprend deux cœur Arm Cortex-A53 tournant à 1 GHz, des codecs vidéo H.265 / H.264 et supporte deux capteurs CMOS[531].
- **Canaan** (2013, Chine) est à l'origine un concepteur de chipsets de mining de bitcoins, derrière son compatriote BitMain. Il est aussi à l'origine de Kendryte K210, un chipset d'inférence basse consommation doté de deux noyaux RISC-V 64 bits et de tenseurs et qui sert notamment à la détection de visage à 60 images par seconde.
- **Kalray** (2008, France, $39,7) est une startup spin-off du CEA qui développe des processeurs « manycore », à 256 cœurs, dédiés au marché de l'automobile et aux applications d'IA associées. La startup annonçait en mai 2018 l'entrée à son capital d'Alliance Ventures, le fonds de capital-risque de l'Alliance Renault-Nissan-Mitsubishi, et de Definvest, le fonds géré par Bpifrance pour le compte du Ministère des Armées.

  Leur environnement de développement en est à sa troisième génération, le KaNN 3.0 (Kalray Neural Network) pour l'adaptation d'applications de deep learning à leur processeur. Ils sont aussi partenaires du projet de processeur européen EPI (European Processor Initiative) qui est piloté par Atos. Kalray fournira la propriété intellectuelle de ses processeurs MPPA (Massively Parallel Processor Array) pour l'associer avec des noyaux Arm généralistes dans le processeur embarqué de l'EPI ciblant le marché automobile.

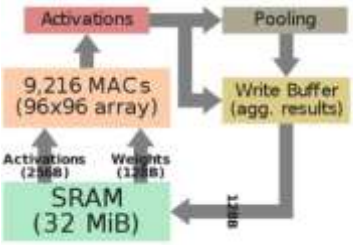
[530] Voir BVS Tech - La caméra biomimétique aux capacités incomparables, Patrick Pirim, mars 2017 (1h43) et Perceptive Invariance and Associative Memory Between Perception and Semantic Representation USER a Universal SEmantic Representation Implemented in a System on Chip (SoC), Patrick Pirim, 2016 (14 pages).

[531] Voir Microprocesseur avec accélérateur d'IA optimisé pour la vision, juin 2020.

- **Tesla** annonçait en avril 2019 lancer son propre chipset, le Full Self-Driving Chip (FSD Chip) pour véhicules autonomes jusqu'au niveau 5 et devenir ainsi autonome vis-à-vis de Nvidia. Une Tesla en comprendrait 2. Fabriqués en 14 nm, ils délivrent chacun 50 TOPS avec 6 milliards de transistors pour environ 70W[532]. Les chipsets comprennent deux NPU avec chacun un gestionnaire de matrices de 96x96 nombres entiers.



Ils supportent directement les fonctions ReLU, sigmoïdes et TanH, couramment utilisées dans les CNNs. Il se distingue en ayant une bonne part de son espace de NPU réservé à 32 Mo de mémoire SRAM pour stocker les hyperparamètres des réseaux de neurones exécutés[533]. Le tout délivre 73,7 TOPS.

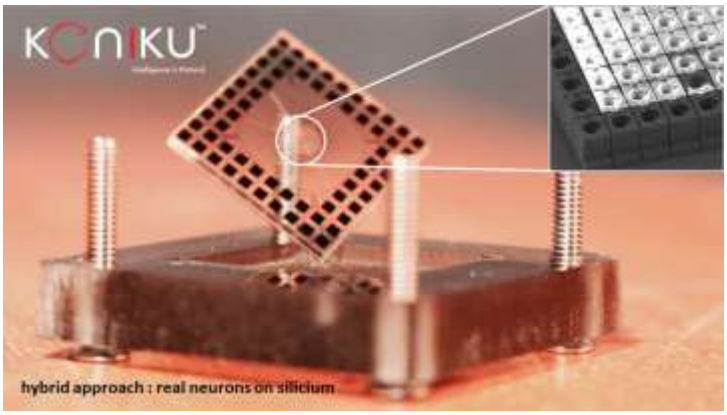


- **Cornami** (2012, USA, $6,5M) propose une architecture de chipset de deep learning massivement parallèle qui peut gérer de plusieurs milliers à millions d'unités de traitement (cores) et facilement gérer des réseaux de neurones notamment convolutifs. Ils appellent cela des matrices systoliques (systolic arrays). C'est associé à une bibliothèque logicielle.
- **Huawei** propose son Ascend 310, un chipset pour le marché de l'embarqué consommant 8 W pour 16 TOPS en entiers 8 bits et 8 TFLOPS en flottant 16 bits. Ce qui nous fait 2 W par TOPS, ce qui est dans la fourchette haute du marché. Il comprend aussi un décodeur de vidéo Full HD.
- **STMicroelectronics** investi aussi le domaine de l'IA dans l'embarqué avec son kit logiciel STM32CubeMX.ai servant à compiler du code de réseau de neurones entraîné et à l'exécuter sur un microcontrôleur classique STM32 à basse consommation. Il supporte les principaux framework open source du marché comme TensorFlow. STMicroelectronics présentait aussi au CES 2018 un prototype de chipset réalisé en 28 nano FD-SOI pour l'exécution directe de réseaux de neurones, comprenant 8 accélérateurs de réseaux convolutifs, 2 DSP et un cœur ARM Cortex M4.
- **SiMa**[ai] (2018, USA, $30M) développe des chipsets d'inférence à basse consommation (10 TOPS/watt et 1000 images par seconde par watt) ciblant divers marchés comme l'automobile, la robotique et la santé.

[532] Voir FSD Chip – Tesla qui détaille l'architecture du chipset. La différence de performance avec l'offre de Nvidia n'est pas si manifeste que cela sur le papier.

[533] A noter qu'il faut 22,7 Mo pour exécuter un réseau convolutionnel ResNet-50 en entiers 8 bits et 62 Mo pour son équivalent avec le SDK YOLOv3 toujours en entiers 8 bits. La gestion de la mémoire dans les NPU est bien expliquée dans AI Inference Memory System Tradeoffs, août 2019.

- **Prophesee** (2013, France, $65,3M) est l'ex Chronocam, chez qui Intel est le plus gros investisseur. Leur chipset est en fait un capteur vidéo qui intègre un réseau de neurones permettant l'interprétation immédiate des images par analyse des différentes d'une image à la suivante. Ils sont partenaires de Sony depuis 2020 avec qui ils ont conçu un capteur CMOS doté d'une grande sensibilité en mode HDR (haute dynamique) [534].
- **Ambarella** (2004, USA, $38,5M+IPO) propose des chipsets de traitement de l'image capables d'exécuter des réseaux de neurones de traitement pour des systèmes embarqués. Leur dernière génération est le CV5, produit en technologie 5 nm, exploitant leur moteur CVflow AI. Il peut supporter des caméras 8K et des systèmes de captation pour la réalité virtuelle ainsi que des systèmes d'aide à la conduite (ADAS).
- **Yumain Sensing** (2011, France), Global Sensing Technologies avant septembre 2019, développe des chipsets de vision artificielle basse consommation exploitant la technologie PNeuro issue du CEA-LIST.
- **CEA-Leti/List** présentait le circuit SamurAI en 2020 qui intègre un accélérateur de machine learning à 64 cœurs, un modem radio à réveil rapide, un noyau RISC-V, un module de chiffrement et la gestion des entrées/sorties pour gérer des capteurs. Il génère 1,3 TOPS/watt et est produit en 28 nm chez STmicroelectronics. C'est le type d'architecture intégrée qui permet de réduire le nombre de composants dans les objets connectés[535].
- Le projet **Nanolitz** aussi financé par la DARPA dans le cadre des projets Atoms to Product (A2P) s'appuie sur des fils microscopiques pour connecter plus efficacement des cœurs et neurones dans des circuits spécialisés.
- Le chipset chinois **Tianjic** imiterait le fonctionnement des neurones du cerveau humain avec 156 cœurs dénommés FCore et 40 000 unités de traitements. Il combine des capacités de réseaux convolutifs pour détecter les objets et de réseaux à mémoire pour gérer le mouvement. Cela lui permet de conduire un vélo robotisé, à même de se déplacer et même d'éviter les obstacles[536].
- **Koniku** (2014, USA, $1,7M) développe des neurones hybrides en silicium et biologiques. La société californienne se positionne aussi sur la reconnaissance d'images. Ils visent les marchés militaires, de la sécurité et de l'agriculture. Mais ils ne donnent pas beaucoup de signes de vie depuis quelques années. Il faut dire que c'était assez space !



[534] Voir Prophesee S.A. and Sony Corporation today announced they have jointly developed a stacked Event-based vision sensor with the industry's smallest 4.86µm pixel size and the industry's highest 124dB (or more) HDR performance, février 2020. Et puis l'annonce produit résultante : Sony dévoile le premier capteur photo dopé à l'IA et met un pied dans "l'edge AI" par Guillaume du Mesgnil d'Engente, mai 2020.

[535] Voir VLSI 2020 : présentation d'un circuit preuve de concept d'intelligence artificielle des objets (AIoT), juin 2020.

[536] Voir Towards artificial general intelligence with hybrid Tianjic chip architecture par Jing Pei & al, 2019 (19 pages) et Voir China's Tianjic Chip Rides A Bike par Mike James, août 2019.

- Des chercheurs du **MIT** présentaient début 2016 leurs travaux sur le chipset **Eyeriss** utilisant des neurones spécialisés répartis dans 168 cœurs dotés de leur propre mémoire. L'application visée est la reconnaissance d'images. Le projet est financé par la DARPA. En 2019, ils en étaient à la version V2[537].

- Pour être complet, il faudrait aussi citer quelques autres startups : **Black Sesame** (2016, USA/Chine, $115M) qui développe des algorithmes et processeurs de vision artificielle ciblant notamment l'assistance à la conduite (ADAS), **DinoPlusAI** (2017, Chine/USA) qui développe un chipset d'inférence à faible latence créé sur une base de FPGA Xilinx Alveo, **FuriosaAI** (Corée du Sud) qui développe un chipset inférence avec un FPGA d'évaluation, avant de faire un ASIC, **TensTorrent** (2016, Canada, $34,5M) également avec des ASIC pour la vision, **Kneron** (2016, Chine, $33M) avec ses ASIC pour la vision, **Corerain Technologies** (2016, Chine), **Esperanto Technologies** (2014, USA, $63M) et ses chipsets basse consommation à base de noyaux RISC-V, **Untether AI** (2017, Canada, $10M) qui optimise le transfert de données dans son processeur, **Maxim Integrated** (USA) et ses micro-contrôleurs arm intégrant **un NPU** à très basse consommation**, Synthiant** (USA, $30M) est ses processeurs à basse consommation, **Aspinity** (2012, USA, $3,5M) et ses chipsets pour objets connectés à basse consommation **RAMP** (**R**econfigurable **A**nalog **M**odular **P**rocessor), **NextVPU** (2016, Chine, $75M) et ses processeurs de vision artificielle, **Perceive** (2016, USA, $1M), une spin-off de Xperi Corp qui développe Ergo, un chipset de 4 TOPS et supportant 55 TOPS/Watt (sans préciser si c'est du 8 bits ou un autre format) fabriqué en 22 nm par Global Foundries, **Realtime Robotics** (2016, USA, $2M) et son chipset optimisé pour le planification de mouvements pour les robots et véhicules autonomes, **aiCTX** (2017, Suisse, $2,8M) et ses **DY**namic **N**euromorphic **A**synchronous **P**rocessors (**DYNAP**s) pour objets connectés, notamment pour de la vision artificielle et de la reconnaissance de gestes, **XMOS** (2005, UK, $94,8M) et ses chipsets de traitement de la parole à base de réseaux de neurones binaires (BNN) et **iniVation** (Suisse) et son Dynamic Vision Sensor qui rappelle la technologie du français Prophesee.

Blocs fonctionnels

Les concepteurs de blocs fonctionnels pour processeurs embarqués qui les commercialisent sous forme de propriété intellectuelle se sont aussi mis à l'IA depuis 2016. Les « blocs d'IP » sont des blocs fonctionnels de chipsets qui sont vendus aux concepteurs de chipsets sous forme de propriété intellectuelle (« blocs de propriété intellectuelle »). Ces blocs fonctionnels sont souvent dédiés soit au traitement de l'image soit à celui du langage en parfois du bruit.

- **Tensilica** (1997, USA, $44M) avec ses Vision C5 surtout destinés aux systèmes de vision artificielle comme les caméras de surveillance. En septembre 2018, Cadence, qui possède Tensilica, annonçait les blocs d'IP Tensilica DNA 100 également adaptés à la vision artificielle, permettant de réaliser des inférences sur plus de 2500 images par seconde[538].

- **arm,** absorbé par Nvidia en octobre 2020, était un peu à la traîne sur l'IA. Il déployait au début une stratégie voisine de celle de Qualcomm avec en premier lieu une offre logicielle dans la plateforme du projet Trillium annoncée en février 2018, supportant les principaux frameworks de deep learning leaders du marché (TensorFlow, Caffee2, Mxnet, Android NNAPI) avec diverses bibliothèques arm les reliant aux blocs fonctionnels d'arm, notamment les Cortex-M utilisés dans de nombreux micro-contrôleurs[539] et les GPU Mali.

---

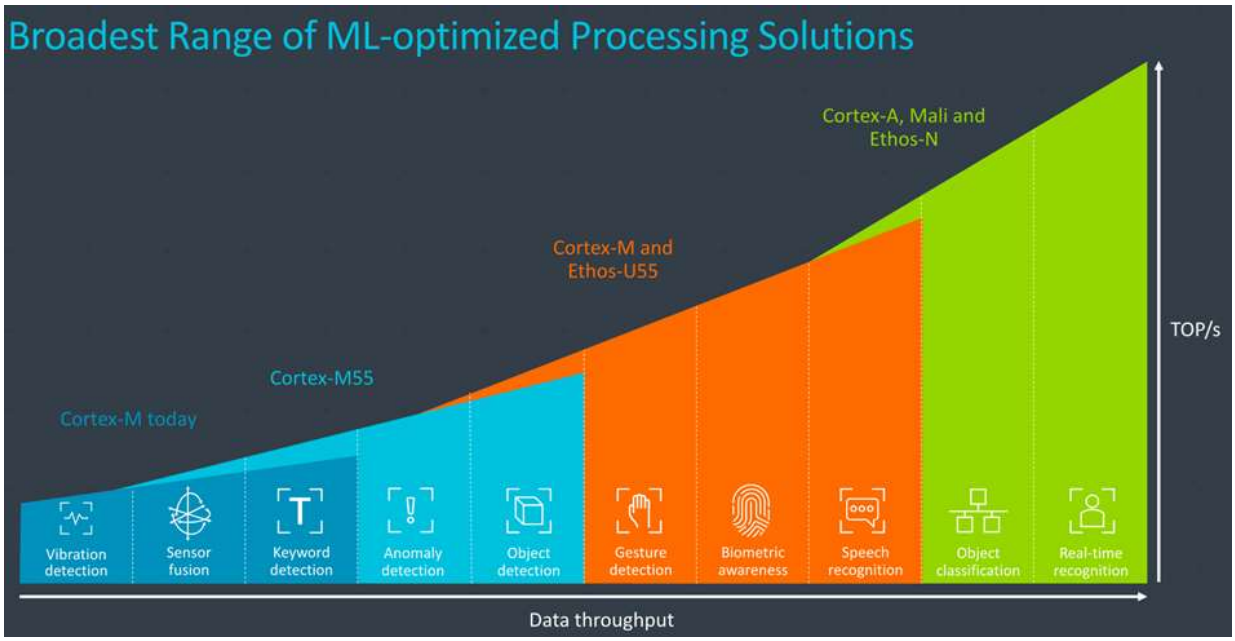
[537] Voir Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices par Yu-Hsin Chen, mai 2019 (21 pages). Ca parle de réseaux maillés hiérarchiques.

[538] Voir Cadence Announces The Tensilica DNA 100 IP: Bigger Artificial Intelligence, septembre 2018.

[539] Voir Machine Learning on Arm Cortex-M Microcontrollers par Naveen Suda, Arm (10 pages).

S'y ajoutait un nouveau bloc d'IP dédié à la détection d'objets dans des réseaux convolutifs[540] avec 4 TOPS, visant 3 TOPS par Watt consommé[541] ce qui ne constitue pas une avancée particulière. En avril 2019, Arm annonçait son nouveau GPU, le Mali G77, gravable en 7 nm, 60% plus rapide que la version antérieure, la G76, pour des tâches d'inférences de réseaux de neurones dans la vision artificielle. Cela reste un GPU classique sans tenseurs et multiplicateurs de matrices et vecteurs. On peut assembler 7 à 16 cœurs G77 dans son chipset.

Cela passait au G78 en avril 2020, pouvant intégrer jusqu'à 28 coeurs. L'adoption de l'IA prenait un tournant en février 2020 avec l'annonce du bloc d'IP Ethos-U55 complétant le nouveau micro-contrôleur Cortex-M55. L'Ethos est positionné comme un « micro-NPU » capable de réaliser du calcul vectoriel et matriciel avec 32, 64, 128 ou 256 modules de MACs (Multiply-Accumulate)[542]. La combinaison Cortex-M55/Ethos est positionnée pour les applications embarquées de détection de gestes et d'objets dans la vision et de reconnaissance de la parole.



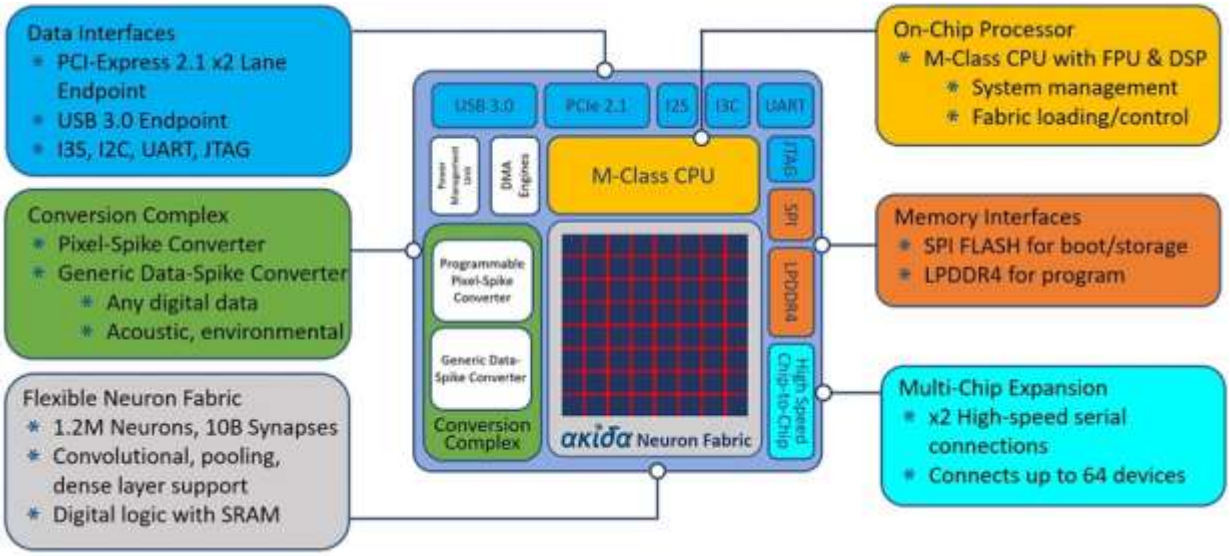
- **eSilicon** (1999, USA, $98M) propose son bloc d'IP neuASIC pour chipsets en technologie ASIC 7 nm pour exécuter des applications de machine learning (détails).
- **AIStorm** (USA, $13,2M) a une approche qui consiste à rapprocher les fonctionnalités de deep learning des sources d'informations dans les capteurs qui génèrent des données analogiques, notamment avec les imageurs CMOS de TowerJazz. Cela en fait visiblement des fournisseurs de propriété intellectuelle de chipsets. Cette approche permet d'améliorer les caractéristiques temps réel des solutions et de baisser la consommation. Ils annoncent un record de 10 TOPS par watt.
- **GreenWaves Technologies** (2014, France, 10M€) est une startup de Grenoble fondée par des anciens de STmicroelectronics. Elle propose le GAP8, un microcontrôleur basse consommation pour les objets connectés générant de 22 à 180 GOPS et réalisé en technologie 22 nm FD-SOI. Il est adapté à l'inférence de réseaux convolutifs pour divers usages : vision artificielle, analyse audio, maintenance préventive, etc. Le chipset contient un accélérateur de calcul convolutif (de type tenseur ?) et huit cœurs de calcul en architecture RISC-V.

[540] Voir Deep Learning on Arm Cortex-M Microcontrollers de Rod Crawford, Arm (28 slides)

[541] Ses fonctionnalités sont détaillées dans Hot Chips 2018: Arm's Machine Learning Core Live Blog, avril 2018.

[542] Voir Arm Announces Cortex-M55 Core And Ethos-U55 microNPU par Andrei Frumusanu, février 2020.

- **Leapmind** (2012, Japon, $13,5M) qui créé des blocs fonctionnels pour réseaux de neurones à basse consommation. Leur solution DeLTA-Lite permet en fait d'entraîner un réseau de neurones et de compiler le modèle résultant pour le graver ensuite dans un composant FPGA.
- **Mipsology** (2015, France) est une startup qui propose Zebra, un framework d'accélération de réseau de neurones qui fonctionne ensuite sur des FPGA, programmés sur des puces de Xylinx et bientôt, d'autres fournisseurs de FPGA. Ces FPGA sont intégrés dans des cartes d'accélération PCI pour serveurs ou dans des composants qui peuvent être intégrés dans des solutions embarquées. Ils supportent les framework Caffe, Caffe2, MXNET et TensorFlow avec des précisions sur 8 ou 16 bits.
- **Brainchip** (2006, Australie, $27,8M) développe des modules de processeurs à neurones par impulsion (spiking neuron adaptive processor ou SNAP) qui sont licenciés à des concepteurs de chipsets. Ils développent aussi les briques logicielles qui exploitent l'architecture. Leurs chipsets Akida comprennent 1,2 million de neurones et 10 milliards de synapses. D'après la littérature de la startup, ils permettent un apprentissage « non supervisé » à la fois pour le traitement du langage et de l'image[543].



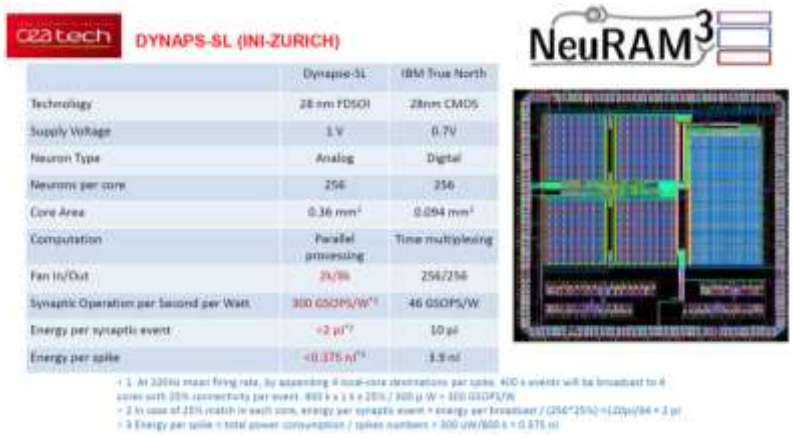
- **Imagination Technologies** (UK) était historiquement un grand acteur des blocs d'IP de GPU, fournissant notamment Apple pour les processeurs de ses premiers iPhone et iPad. Depuis qu'Apple les a abandonnés pour créer ses propres GPU, Imagination Technologies a dû faire un pivot. Ils s'intéressent maintenant principalement au marché autonomobile et créent des blocs de réseaux de neurones basse-consommation.
- **Eta Compute** (2015, USA, $8M) propose Tensai, une puce neuromorphique censée être capable de faire de l'apprentissage de manière autonome. Elle exploite leur architecture à logique asynchrone DIAL (Delay Insensitive Asynchronous Logic) qui permet d'utiliser de très basses tensions, ce qui réduit d'autant la consommation d'énergie. L'ensemble est adapté aux réseaux convolutifs (CNNs) et à impulsions (SNNs) pour divers objets connectés. La puce exploite en fait des noyaux combinant un Arm Cortex-M3 et un NXP CoolFlux DSP16. Un noyau simule 2000 neurones et 128 000 synapses[544].

---

[543] Source du schéma : https://www.brainchipinc.com/products/akida-neuromorphic-system-on-chip. Et la présentation du lancement d'Akida : Brainchip Akida Launch Presentation, 2018 (27 slides). Voir Neuromorphic Chip Maker Takes Aim At The Edge par Michael Feldman, janvier 2020.

[544] Leur architecture est décrite dans On the Link Between Energy and Information for the Design of Neuromorphic Systems par Narayan Srinivasa, avril 2018 (19 slides).

- **Flex Logix Technologies** (2014, USA, $12,4M) propose des blocs d'IP de gestion de réseaux de neurones, en particulier pour la reconnaissance d'images, et destinés à des FPGA intégrables dans des cartes PCI. Ils ciblent de nombreux marchés : robots, caméras de surveillance, automobile, set-top-boxes et même serveurs[545]. Le tout est supporté par TensorFlow Lite et ONNX.
- Côté recherche appliquée, le **CEA-Leti** a lancé le projet de chipsets neuromorphiques N2-D2 en technologie FD-SOI 28 nm couvre les besoins des systèmes embarqués et de la reconnaissance d'images. Le CEA est aussi impliqué dans la conception de la puce basse consommation Dynapse-SL du projet H2020 européen Neuram3 impliquant aussi la Suisse et IBM[546].



Neurones à impulsions

Une autre piste est explorée qui consiste à imiter d'encore plus près le fonctionnement des neurones biologiques avec les systèmes à base de neurones à impulsions, ou *spiking neurons*[547]. Celles-ci s'appuient aussi sur des unités de calcul associant traitement et mémoire. L'un de ses intérêts est d'être très économe en énergie[548].

Le neurone reçoit un train d'impulsions dans l'ensemble de ses synapses et génère en sortie un train d'impulsions, le résultat du calcul. Ces *spiking neurons* doivent être programmées de manière spécifique[549]. Par contre, leur inconvénient est qu'il est difficile de les entraîner[550]. L'un des enjeux est aussi d'améliorer la densité des neurones[551].

C'est la voie choisie par **IBM** avec ses chipsets TrueNorth ainsi que par **Intel** avec ses chipsets Loihi puis **SpiNNaker** et **BrainScales** qui ont été créés dans le cadre du projet européen Human Brain Project, et enfin le Français **GrAI Matter Labs** déjà cité plus haut.

---

[545] Voir AI On The Edge: New Flex Logix X1 Inference AI Chip For Fanless Designs, avril 2019.

[546] Quelle différence entre ces deux projets? N2-D2 est un chipset neuromorphique classique qui intègre la logique de traitements. Dynapse-SL est un chipset qui intègre la logique et la mémoire, avec des neurones analogiques, permettant d'améliorer les performances et de réduire encore plus la consommation d'énergie. Le CEA-Leti bosse aussi sur des chipsets avec des neurones à impulsion, dans le cadre du projet Spirit et exploitant de la mémoire OxRAM. Voir Neuromorphic and Deep Learning Technologies at CEA, Marc Duranton et Carlo Reita, 2017 (57 slides), NeuRAM3: NEUral computing aRchitectures in Advanced Monolithic 3D-VLSI nano-technologies, 2017 (7 slides) et OxRAM Memories A Disruptive Technology For Disruptive Designs, 2017 (30 slides).

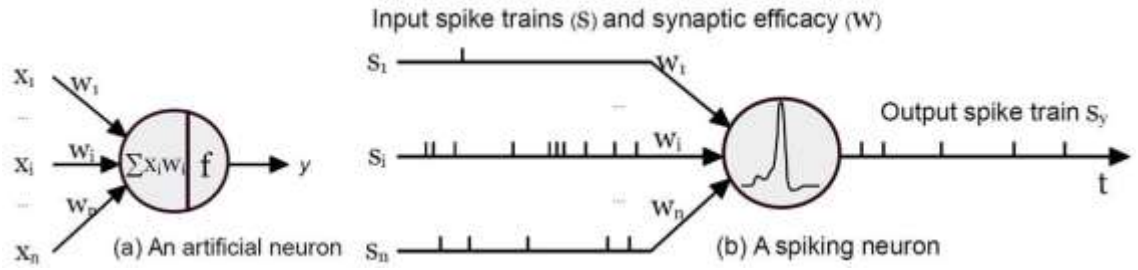
[547] Cette thèse Deep Spiking Neural Networks de Qian Liu, 2018 (212 pages) fait un point à jour de la technologie des spiking neurons. Elle porte sur leurs techniques d'entraînement, offline (sur d'autres architectures) et online (dans les spiking neurons). C'est de cette thèse que provient l'illustration de cette page avec la comparaison entre neurone artificielle classique et neurone à impulsion. Voir aussi Low-Power Neuromorphic Hardware for Signal Processing Applications par Bipin Rajendra & Al, août 2019 (24 pages).

[548] Pour diminuer encore plus la consommation énergétique, des chercheurs imaginent même utiliser des composants supraconducteurs, devant donc être refroidis à très basse température pour fonctionner. Voir Design of a Power Efficient Artificial Neuron Using Superconducting Nanowires par Emily Toomey et al, septembre 2019.

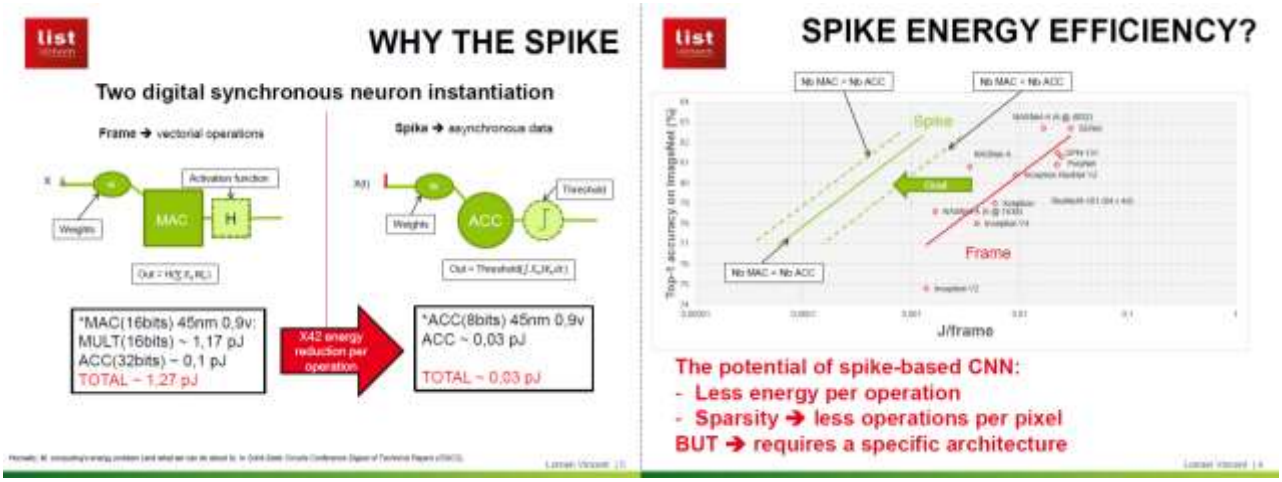
[549] Voir par exemple Project Ihmehimmeli: Temporal Coding in Spiking Neural Networks par Iulia-Maria Comșa et Krzysztof Potempa de Google Research, Zürich, septembre 2019.

[550] Voir Deep Learning With Spiking Neurons: Opportunities and Challenges par Michael Pfeiffer et Thomas Pfeil, 2020.

[551] Des chercheurs indiens ont trouvé une méthode pour aller dans ce sens comme décrit dans New Hardware Mimics Spiking Behavior of Neurons With High Efficiency par Michelle Hampson, juin 2020, qui fait référence à Band-to-Band Tunneling based Ultra-Energy Efficient Silicon Neuron par Tanmay Chavan et al, 2019 (20 pages).

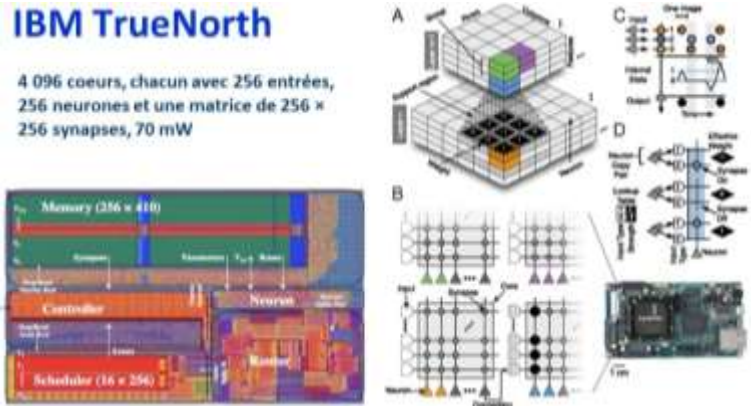


Le projet **NeuroSpike** du CEA porte sur la création de puces à neurones à impulsion à très basse consommation. Le gain de consommation par rapport à des puces matricielles classiques serait de x42 (schéma *ci-dessous*). Utilisable dans les CNN. Le système protégé par plusieurs brevets consomme 11 fois moins d'énergie et est 4 fois plus rapide pour l'entraînement que les chipsets matriciels.



IBM TrueNorth a été développé dans le cadre du programme SyNAPSE de la DARPA. Lancé en 2014, les chipsets TrueNorth sont capables de simuler un million de neurones artificiels, 256 millions de synapses reliant ces neurones et exécutant 46 milliards d'opérations synaptiques par secondes et par Watt consommé. Le tout avec 4096 cœurs fonctionnant par multiplexage temporel et 5,4 milliards de transistors, donc aux alentours d'une vingtaine de transistors par synapses. Il faut noter que dans le cas de TrueNorth, ces neurones ne peuvent pas être exploités dans un processus d'apprentissage. Leur entraînement doit être réalisé sur d'autres machines.

Le chipset initial a été fabriqué par Samsung en technologie CMOS 28 nm et avec une couche d'isolation d'oxyde de silicium (SOI, avec des wafers issus du français **SOITEC**) permettant de diminuer la consommation électrique et d'accélérer les traitements. Le chipset fait plus de 4 cm^2 de surface et ne consomme que 70 mW, ce qui permet d'envisager d'empiler ces processeurs en couches.



C'est quelque chose de difficile à réaliser avec les processeurs CMOS habituels qui consomment beaucoup plus d'énergie au cm^2.

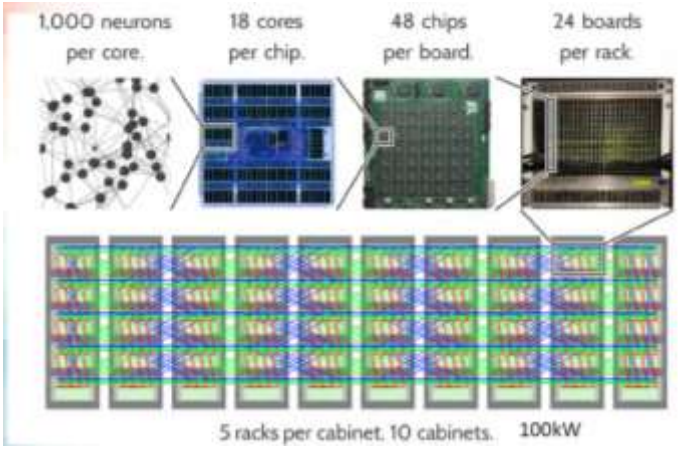
À titre de comparaison, un processeur Intel Core i7 de génération Coffee Lake réalisé en technologie 14 nm consomme entre 15 W et 130 W selon les modèles, avec 5 milliards de transistors et un GPU Nvidia V100 de 21 milliards de transistors consomme 300 W. Le but d'IBM est de construire un ordinateur doté de 10 milliards de neurones et 100 trillions de synapses.

Il consommerait 1 KW et tiendrait dans un volume de deux litres.

Une étape intermédiaire a été annoncée au printemps 2017 : un ordinateur neuromorphique développé pour l'US Air Force et doté de 64 millions de neurones mais dont le domaine d'application n'a pas été précisé.

Mais cette première version de TrueNorth n'avait pas de mémoire intégrée. Il en va autrement d'une nouvelle mouture de chipset neuromorphique IBM testée en 2018 qui intègre des memristors PCM (Phase Change Memory) [552] . Cela a permis d'améliorer encore l'efficacité énergétique de l'ensemble, mais avec seulement 204 900 synapses pour commencer[553]. Les tests ont été réalisés sur un algorithme de reconnaissance d'écriture appliqué sur la base standard MNIST. Ce chipset réalisait 29 milliards d'opérations par seconde et par Watt consommé, un peu moins que le chipset TrueNorth d'origine.

SpiNNaker est un autre processeur à neurones à impulsion, créé dans le cadre du projet européen Human Brain Project par Steve Furber (Université de Manchester, UK). Il vise à simuler le fonctionnement d'un milliard de neurones.



Il s'appuie sur une architecture matérielle avec 18 cœurs 32 bits ARM de 18 000 neurones par chipset. On est plus dans l'architecture massivement parallèle avec des milliers de processeurs de ce type que dans les processeurs véritablement synaptiques.

L'architecture est suffisamment souple pour exécuter différents types de réseaux de neurones y compris des neurones à impulsions. C'est un projet de recherche. Toujours au sein du HBP, le projet **BrainScales** de 2010 à 2015 s'appuyait sur un chipset de 112 650 synapses programmables et 512 neurones fonctionnant par impulsions[554].

---

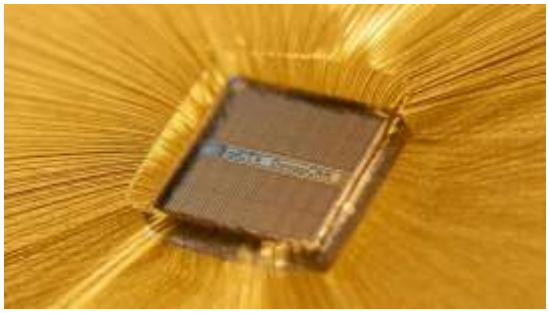
[552] Voir Reliability enhancement of phase change memory for neuromorphic applications de SangBum Kim, IBM, 2017 (21 slides).

[553] Voir AI could get 100 times more energy-efficient with IBM's new artificial synapses, juin 2018 et Equivalent-accuracy accelerated neural-network training using analogue memory, 2018. IBM communique sur le fait qu'ils utilisent une mémoire analogique. C'est en effet le cas pour les informations stockées dans les memristors.

[554] Voir Neuromorphic Computing in the HBP par Steve Furber de l'Université de Manchester et Karlheinz Meier de celle d'Heidelberg en Allemagne, 2017 (42 slides).

GrAI Matter Labs (2016, France, $29M) conçoit un processeur sur une architecture à base de réseaux neuronaux, numérique mais asynchrone, utilisant des « spiking neurons » (neurones à impulsion). Ils ambitionnent d'intégrer un million de neurones sur un centimètre carré et consommant 1 W, programmable en Python. Ils ont obtenu $1M de financement de la DARPA pour un démonstrateur FPGA qui tourne déjà. C'est rare pour une startup française. Les marchés visés sont les véhicules autonomes, la maison connectée et la santé.

SynSense (2017, Suisse, $2,8M), anciennement aiCTX, développe des solutions matérielles et logicielles neuromorphiques en valorisant les travaux de l'Institut de neuroinformatique de l'ETH Zurich. DynapCNN gère un million de neurones à impulsion et quatre millions de paramètres adaptée aux applications embarquées à très basse consommation. Elle sert notamment aux applications de vision artificielle pour la détection d'événements en temps réel et avec une très faible latence de 5 ms.



Il supporte tous types de réseaux de neurones convolutifs. La puce ASIC est fabriquée en 22 nm et fait seulement 12 mm^2. Ils proposent aussi le DynapSEL qui est adapté à divers réseaux de neurones, notamment récurrents ou à réservoirs et aux applications dans la santé et la robotique. Il comprend un millier de neurones à impulsion et 80 000 synapses configurables.

Reexen Technology (2017, Suisse et Shenzhen) développe des chipsets d'accélération intégrés dans une solution complète avec logiciels. La société a été lancée par des anciens de l'ETH Zurich. Cela semble un concurrent de aiCT, qui est basé aussi à Zurich.

Enfin, **Neurocore** est un projet de l'Université de Stanford, aussi associé au HBP, un chipset intégrant 65 536 neurones et fonctionnant à très basse consommation.

Memristors

Les memristors ont été conceptualisés en 1971 par le sino-américain **Leon Chua**[555]. Ce sont des composants électroniques capables de mémoriser un état en faisant varier une résistance électrique par l'application d'une tension, un peu comme les cristaux liquides bistables qui servent dans les liseuses électroniques. La valeur modifiable de la résistance permet de stocker de l'information et de manière non volatile. C'est particulièrement utile pour créer des réseaux de neurones avec des poids de synapses gérés dans ces résistances.

Ces poids peuvent être lus et modifiés rapidement sans avoir besoin d'accéder à une mémoire externe, qui bien est lente par rapport à la vitesse de fonctionnement du processeur, en particulier dans les phases laborieuses d'entraînement du deep learning. Dans certains cas de figure, cela permet même de réaliser un entraînement temps réel incrémental de réseau de neurones. C'est un type d'architecture de chipsets envisagée depuis des années mais difficile à réaliser[556].

Les memristors peuvent aussi servir à la partie logique des composants et remplacer des transistors classiques ou bien être intégrés au côté de transistors actifs classiques dans des unités de traitement[557]. Dans le premier cas, cela pourrait aboutir à une nouvelle génération de circuits FPGA dont la logique est dynamiquement programmable. Le second cas n'est pas encore opérationnel.

---

[555] Voir Memristor-The Missing Circuit Element de Leon Chua, 1971 (13 pages).

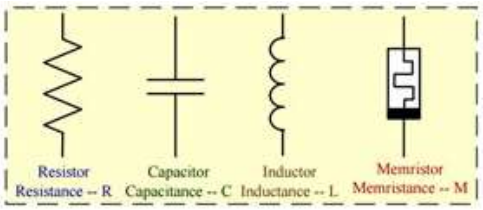
[556] Voir The Search for Alternative Computational Paradigms de Naresh Shanbhag, 2008 (11 pages) qui décrit diverses architectures rapprochant calcul et mémoire. Ainsi que Cognitive Computing at the Limits du même auteur, 2017 (19 slides).

[557] C'est très bien expliqué dans Memristor: From Basics to Deployment de Saraju Mohanty, publié en 2013 (13 pages) et dans Hybrid CMOS/Memristor Circuits 2010 (4 pages).

Par contre, les memristors sont difficiles à mettre au point car ils ne restent pas stables longtemps et leur temps de commutation est parfois trop long. Ils sont aussi difficiles à miniaturiser au même niveau que les mémoires volatiles et non volatiles actuelles.

Les memristors sont produits avec des procédés de fabrication voisins de ceux du CMOS, en ajoutant une bonne douzaine d'étapes dans la production avec des procédés spécifiques de dépôt sous vide de couches minces de matériaux semi-conducteurs.

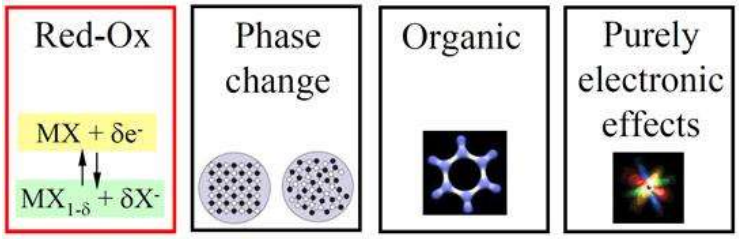
Les fabs investies sur le sujet sont notamment celles de Global Foundries, TSMC et Samsung.



Placer la mémoire à proximité des neurones permet aussi d'économiser de l'énergie et de se rapprocher un tant soit peu de l'efficacité énergétique extraordinaire du cerveau humain, qui ne consomme que 20 Watts.

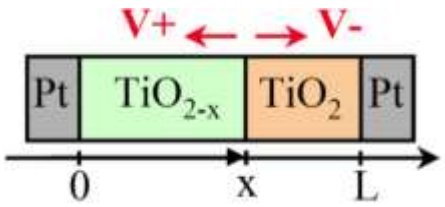
En plus des réseaux de neurones, les memristors ont évidemment comme domaine d'application les mémoires non volatiles pour créer des systèmes de stockage de type SSD. Le défi est de descendre en-dessous de 15 nm pour la taille de chaque cellule mémoire et de supporter un grand nombre de cycles d'écriture et de lecture et une longue durée de vie du système de stockage.

Il existe différents types de memristors qui ont chacun leurs avantages et inconvénients. La recherche progresse depuis près de 20 ans dans le domaine, mais on n'est malheureusement pas encore au stade de l'industrialisation.

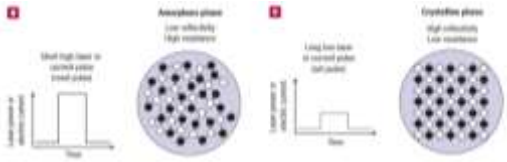


On compte ainsi :

- Les **Red-Ox**[558], des résistances variables par migration d'ions d'oxyde de titane, de tantale, d'argent ou de silicium[559], par procédé de réduction-oxydation. Les ReRAM annoncées par HP en 2013 en font partie[560]. Depuis, HP travaille sur leur fiabilité, leur vitesse de commutation et leur endurance[561].



- Les memristors à **changement de phase** ou PCM (Phase-Change Memory) qui alterne phase amorphe et phase cristalline pour des composites comme du verre de chalcogénure GST (germanium-antimoine-tellure)[562].



[558] Voir Memristive devices for computing de Joshua Yang, Dmitri Strukov et Duncan Stewart, 2013 (12 pages). Plus généralement, voir RRAM for Future Memory and Computing Applications par Ming Liu, juillet 2018 (42 slides).

[559] Weebit Nano est une startup lancée en 2015 en Australie qui développe de la ReRAM à base d'oxyde de silicium. Le SiOx est utilisé dans les transistors. La startup a démontré en juin 2018 une première puce de 1 mbits en technologie 40 nm. La démonstration avait lieu au CEA-Leti à Grenoble, leur partenaire.

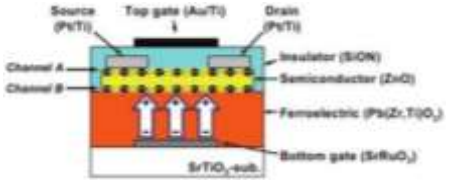
[560] Ces memristors se sont retrouvés dans le HP Dot Product Engine, un processeur expérimental destiné au calcul matriciel à base d'oxyde de tantale. Voir Dot-Product Engine for Neuromorphic Computing: Programming 1T1M Crossbar to Accelerate Matrix-Vector Multiplication par Miao Hu et al, 2016 (7 pages).

[561] Les schémas sont issus de la présentation Neuromorphic computing - Memristors de Julie Grollier, du CNRS / Thalès TRT, 2018 (80 slides). Voir aussi dans son blog : http://julie.grollier.free.fr/memristors.htm. Le Coréen SK Hynix met au point de son côté une ReRAM en technologie 23 nm à base se silicium dopé à l'arsenic.

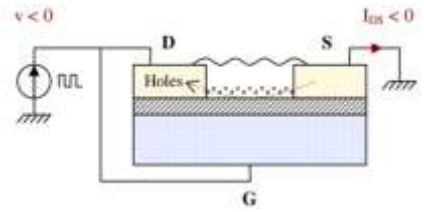
[562] Cette technologie a été créée par une chercheuse Turco-Américaine, Duygu Kuzum.

Les chipsets pour SSD 3D Xpoint d'Intel utilisent une variante de ce type de memristors. On y range les PRAM et les PCRAM.

- Les memristors **ferroélectriques** qui exploitent un changement de polarité par champ électrique (source du schéma). La technique exploite de l'or, du plomb, du zirconium et du ruthenium qui ne sont pas des éléments très abondants.



- Les memristors **organiques**[563], avec des polymères divers comme les NOMFET (Nanoparticle Organic Memory Field-Effect Transistor) avec de l'or et du pentacène, un hydrocarbure à base d'hydrogène et de carbone, des nanotubes de carbone (CNT-FET) et des polyanilines (PANI)[564].



- Les memristors à base de **spintronique** qui sont utilisables dans des MRAM (Magnetic RAM) avec la modification d'un état magnétique qui affecte une résistance, via un isolant entre deux métaux ferromagnétiques.

L'un des problèmes à résoudre est la durabilité des memristors, notamment pour ceux qui exploitent des techniques qui déplacent des atomes. Ceux-ci claquent au bout d'un million de cycles du fait de failles cristallines. C'est moins le cas avec les memristors à changements de phase.

On peut aussi éviter les mouvements d'atomes en jouant sur des jonctions à effet tunnel et à mouvements d'électrons. Dans le cas de memristors utilisés pour des poids synaptiques, il faut aussi réduire le bruit et les erreurs qui génèrent des poids imprécis. L'autre solution demanderait d'exploiter des réseaux de neurones à synapses binaires, avec seulement deux valeurs comme dans les perceptrons d'origine.

Yoshua Bengio de l'Université de Montréal travaille là-dessus, avec les **Quantized Neural Networks**[565] (QNN) qui auraient des performances voisines vis-à-vis des réseaux de neurones classiques, comme pour la reconnaissance d'images dans le jeu de test ImageNet.

Des memristors ont été notamment développés dans le cadre des projets de recherche du programme **SyNAPSE** de la DARPA (2008 à 2016).

HPE a été le premier à en prototyper en 2008 en oxyde de titane et lancé un partenariat avec le fabricant de mémoires coréen **Hynix**, mais le projet a été mis en veilleuse en 2012. Les fabricants de mémoire sont occupés par les DRAM et les mémoires Flash intégrées dans les SSD. Il est difficile pour l'instant de justifier économiquement le lancement d'une fab dédiée aux memristors.

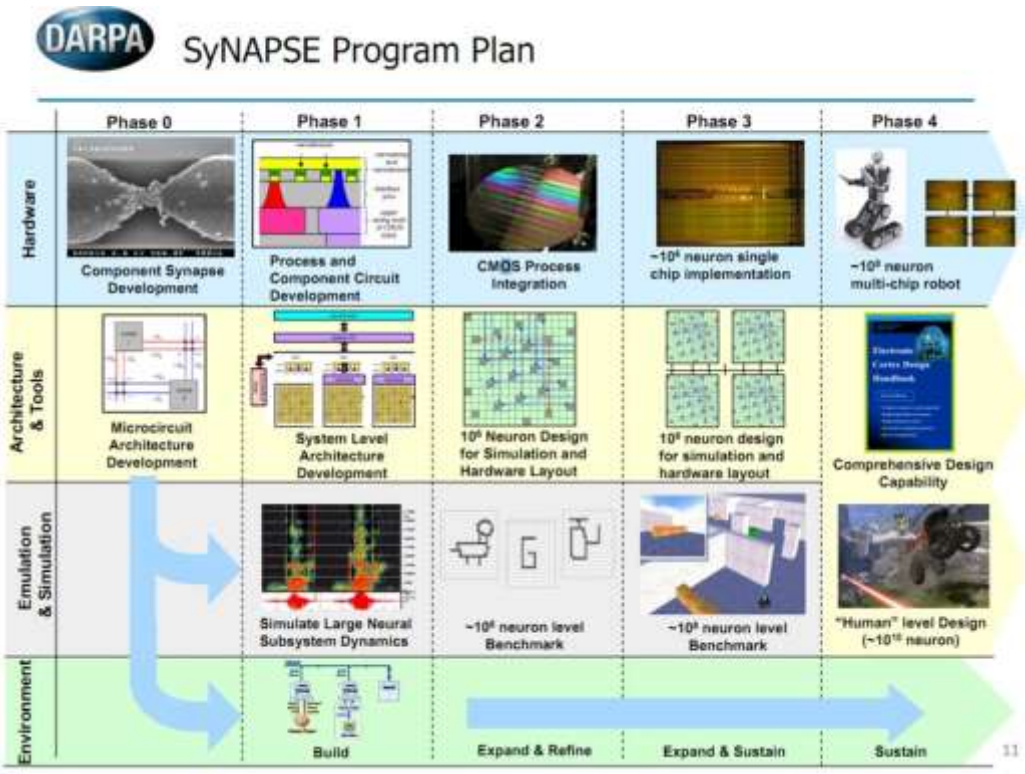
Sinon, le taux de rebut de la fabrication expérimentale était trop élevé, un paramètre clé pour pouvoir fabriquer des composants en quantité industrielle et à un prix de vente abordable. De plus, le nombre de cycles d'écriture semblait limité pour des raisons chimiques, dans le cycle de libération/captation d'oxygène pour les memristors en oxydes de titane.

---

[563] Voir Organic memristors come of age de Ilia Valov et Michael Kozicki, 2017 (2 pages).

[564] Une équipe de chercheurs de l'Université de Singapour menée par Thirumalai Venky Venkatesan met au point des memristors capable de changer d'état en 50 ns et restant stable pendant 11 jours sans alimentation. Ils sont réalisés à base de ruthenium relies par des molécules organiques azotées conçues à l'Université de Yale et réalisées en Inde. Voir Organic Memristor Sets Records for Speed and Durability, octobre 2017.

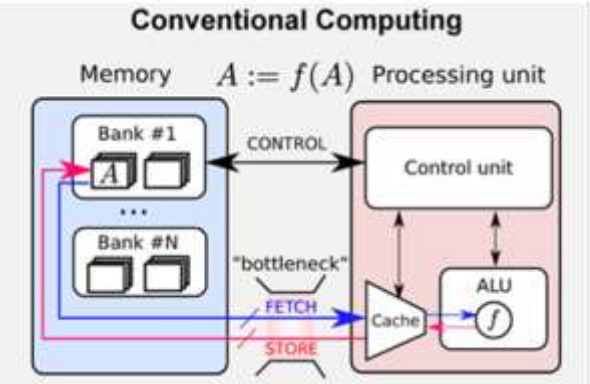
[565] Voir Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations, 2018 (30 pages).

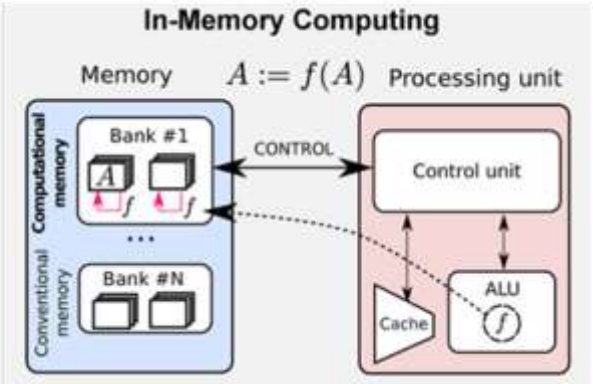


En octobre 2015, HP et **SanDisk** annoncaient un partenariat pour fabriquer des mémoires volatiles et non volatiles à base de memristors, censées être 1000 fois plus rapides et plus endurantes que les mémoires flash traditionnelles. Quatre ans plus tard, il est difficile de savoir ce qu'il en est advenu.

D'autres laboratoires de recherche et industriels planchent aussi sur les memristors :

- Des chercheurs d'**IBM Research Zurich** expliquent bien l'intérêt de déplacer en mémoire certaines fonctions de calcul de réseau de neurones pour les accélérer. On appelle cela le "in memory computing". Les calculs sont réalisés dans des memristors et donc de manière analogique, ce qui peut générer des problèmes de précision[566].



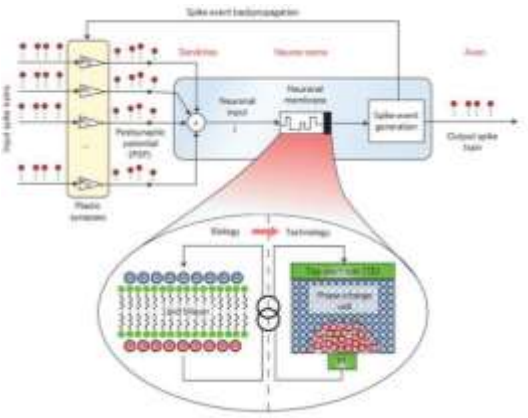


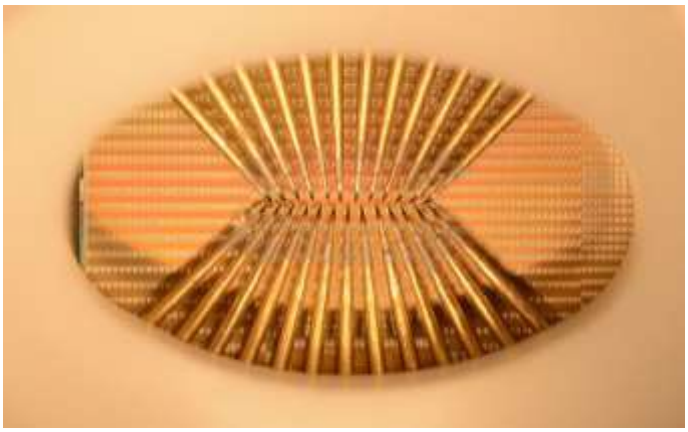
- Des chercheurs de Stanford, San Jose, Singapour et du CEA-Leti à Grenoble présentaient début 2021 leur architecture **Illusion** de multiprocesseur de traitement de réseau de neurones exploitant huit puces intégrant chacune plusieurs puces de mémoire résistive (RRAM) issue du Leti et assemblées en 3D entre elles et avec la puce assurant le calcul. Le tout avec une excellente efficacité énergétique voisine à quelques % près d'un processeur unique à mémoire intégrée[567].

[566] Voir In-Memory Computing: Towards Energy-Efficient Artificial Intelligence par Manuel Le Gallo, Abu Sebastian et Evangelos Eleftheriou d'IBM Research Zurich, octobre 2018.

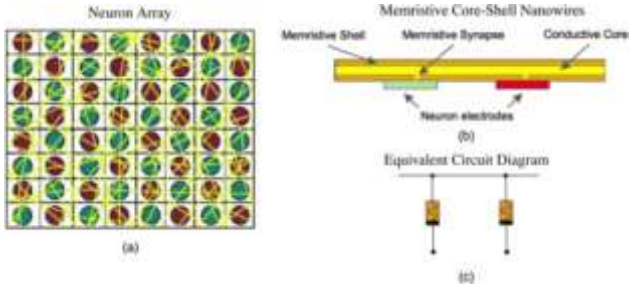
[567] Voir System Creates the Illusion of an Ideal AI Chip par Samuel K. Moore, janvier 2021.

- L'**ETH** de Zurich (le CNRS suisse) développe un memristor capable de stocker trois états à base de pérovskite (titanate de calcium) de 5 nm d'épaisseur[568]. Cela pourrait servir à gérer de la logique floue. Ils explorent aussi la piste des réseaux de neurones[569] à base de memristors à changement de phase en GST (germanium-antimoine-tellure).





- Des chercheurs de l'**Université Technologique du Michigan** annoncaient début 2016 avoir créé des memristors à base de bisulfite de molybdène qui ont un comportement plus linéaire[570]. Cela permettrait de créer des memristors de meilleure qualité et stabilité.
- La startup californienne **Knowm** (2002, USA) commercialise depuis 2016 un composant commercial à base de memristors, fabriqué en partenariat avec la Boise State University, à base d'argent ou de cuivre et au prix de $220. Il est destiné en premier lieu aux laboratoires de recherche en réseaux neuronaux.
- Une équipe de chercheurs associant le **CNRS** et **Thales** située à Palaiseau et pilotée par Julie Grollier travaille sur une technologie avancée de processeurs neuromorphiques, en collaboration avec des laboratoires de recherche japonais (AIST Tsukuba) et américains[571]. Ils utilisent des oscillateurs de résonance magnétique qui permettent de se rapprocher encore plus du mode de fonctionnement des neurones biologiques en facilitant la propagation temporelle des valeurs entre les neurones d'un système, avec des fréquences allant de 300 MHz à 10 GHz, exploitable dans les réseaux de neurones récurrents qui font de la reconnaissance de la parole. Cela permettrait au passage de créer des nano neurones faisant quelques dizaines de nanomètres de large.
- **Rain Neuromorphics** (2018, USA, $2M) est une startup qui met au point une architecture de chipset "Memristive Nanowire Neural Network"[572] qui serait rapide, puissante et très scalable. Les neurones sont connectés entre eux par des nanofils (nanowires) disposés un peu aléatoirement pour relier les neurones entre eux.



[568] Voir Swiss researchers have created a memristor with three stable resistive states, 2015.

[569] Voir IBM creates world's first artificial phase-change neurons par Sebastien Anthony, 2016.

[570] Voir Molybdenum disulfide memristors; neural network chip for mobile; nanoscale materials for the IoT, de Jesse Allen, février 2016.

[571] Voir Neuromorphic computing with nanoscale spintronic oscillators, janvier 2017 (13 pages).

[572] Ils sont documentés dans la présentation Memristive Nanowire Neural Networks (17 slides) par Jack Kendall, CTO de la startup, et par Juan Nino qui enseigne la physique des matériaux à l'Université de Floride.

Et cela utilise du deep learning avec la technique du “reservoir network”. C’est une approche originale qui mérite le détour.

- L’**ANR** française a financé le projet collaboratif **MHANN** associant Inria, l’IMS de Bordeaux et Thalès pour créer des memristors ferromagnétiques. Le projet devait être terminé en 2013 et avait bénéficié d’une enveloppe de 740 K€.
- L’**Université de Tohoku** au Japon annonçait en 2019 avoir conçu une mémoire MRAM de 128 Mbits en technologie STT-MRAM (spin-transfer torque magnetoresistive random access memory) avec une vitesse d’écriture record de 14 ns[573]. Elle serait adaptée à la conception de NPU intégrables dans des chipsets pour les marchés de l’embarqué.

Ce petit tour des memristors montre que l’IA a de la marge d’amélioration si le matériel suit. Il suivra à son rythme, mais une chose est sûre : le rapprochement de la mémoire des zones de traitement dans les réseaux de neurones est un des points de passage obligés pour décupler leur performance.

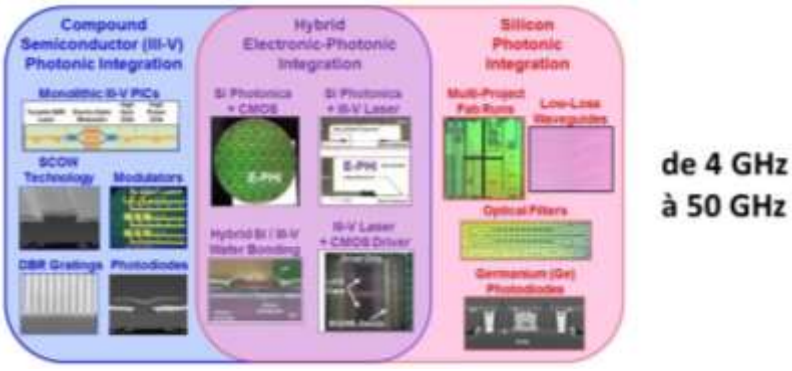
Photonique

La photonique utilise la lumière et les photons pour gérer ou transmettre l’information. Aujourd’hui, dans l’informatique, la photonique est principalement utilisée dans la transmission d’informations sur fibres optiques. Des systèmes de multiplexage permettent de transmettre d’énormes volumes d’information sur une simple fibre optique, jusqu’à des teraoctets par seconde. Elle exploite des composants à base des matériaux dits “III-V” qui associent des éléments correspondant à deux colonnes du tableau de Mendeleev.

Aujourd’hui, la photonique est surtout utilisée dans le multiplexage de données sur les liaisons ultra-haut-débit des opérateurs télécoms, dans des applications très spécifiques, ainsi que sur des bus de données optiques de supercalculateurs. L’intérêt de la photonique est de permettre d’utiliser des fréquences d’horloge énormes, allant jusqu’à 50 GHz voire même au-delà. L’un des enjeux se situe dans l’intégration de composants hybrides, ajoutant des briques en photonique au-dessus de composants CMOS plus lents. Intel et quelques autres sont sur le pont.

Une fois que l’on aura des processeurs optiques généralistes, il faudra relancer le processus d’intégration. Il est actuellement situé aux alentours de 200 nm pour la photonique et la course se déclenchera alors pour descendre vers 10 à 5 nm comme pour le CMOS actuel. Mais un grand nombre d’obstacles technologiques sont à franchir, ne serait-ce que pour créer une logique de traitements en exploitant des photons au lieu d’électrons, ce qui n’est pas une mince affaire[574].

Il existe bien des tentatives de créer des réseaux de neurones en photonique, mais le remplacement de transistors en silicium et technologie CMOS par des transistors en III-V gérant des photons transitant via des fibres optiques est loin d’être facile. Les mécanismes de circulation des photons ne sont pas les mêmes que ceux des électrons !

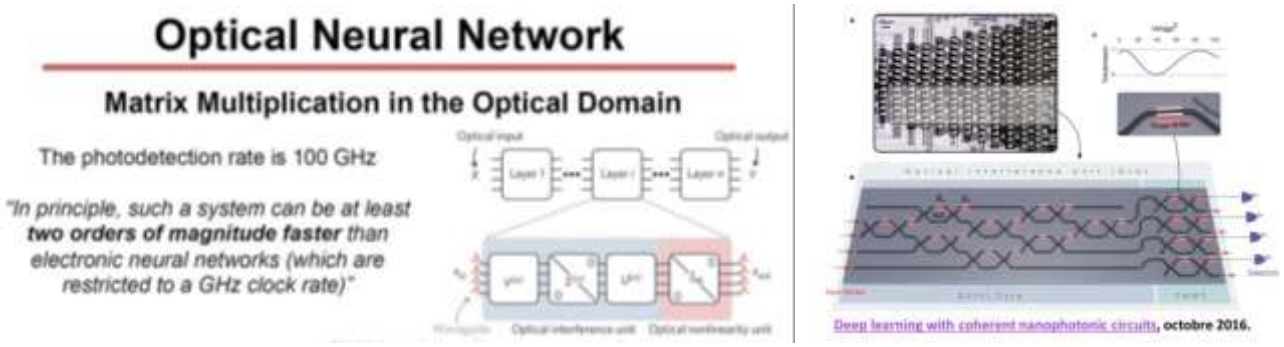


[573] Voir Researchers develop 128Mb STT-MRAM with world's fastest write speed for embedded memory, décembre 2018.

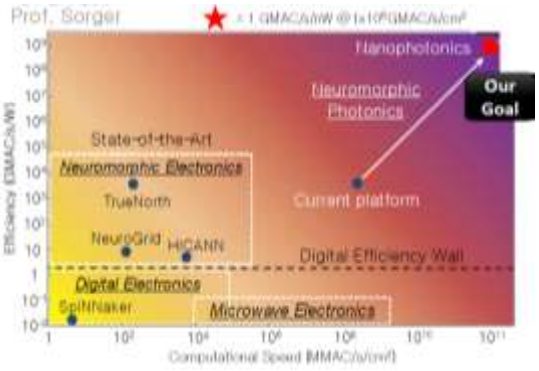
[574] Voir ce reviw paper sur le sujet : Photonics for artificial intelligence and neuromorphic computing par Bhavin J. Shastr et al, novembre 2020 (21 pages).

Les premiers chipsets expérimentaux photoniques à réseau de neurones ont fait leur apparition en laboratoire, avec à la clé un potentiel de multiplication de la performance par 50[575] !

Lightmatter (2017, USA, $30M) et **Lightelligence** (2017, USA, $10M) sont des spin-offs du projet du MIT qui en est à l'origine et qui testent ces chipsets optiques pour faire de la reconnaissance de la parole, mais seulement pour l'exécution et pas pour l'entraînement[576].



L'équipe de Volker Sorger de **GWU** (Georges Washington University) met aussi au point des réseaux de neurones convolutifs en optronique utilisant des composants réalisant des transformées de Fourier rapides[577]. Elle affiche des gains de performance théoriques vertigineux de 10^7 en consommation d'énergie par rapport aux processeurs TrueNorth d'IBM et de 10^9 en rapidité par rapport aux GPU Nvidia (*schémas suivants*) !



Deep Learning

Reference	Efficiency (J/MAC)	Speed (MAC/s)
NVIDIA GPU [15]	3.4×10^{-7}	1.7×10^{7}
AlexNet FPGA [16]	2.6×10^{-10}	6.2×10^{10}
UPSIDE Crossbar [17]	1.3×10^{-10}	4.0×10^{10}
UTK Analog Engine [18]	1.0×10^{-12}	11.8×10^{6}
IBM TrueNorth [19]	2.6×10^{-11}	1.3×10^{6}
Nanophotonic Neuromorphic	$\mathbf{7.4 \times 10^{-18}}$	$\mathbf{2.0 \times 10^{17}}$

IBM Zurich travaille aussi dans ce domaine et développe des réseaux de neurones matriciels optiques qui stockent les poids de neurones sous forme d'indices de réfraction dans des réseaux optiques[578].

---

[575] Voir Deep learning with coherent nanophotonic circuits de Yichen Shen et al, 2017 (7 pages), cette autre approche de traitement et de stockage optique de l'information réalisée par un laboratoire australien : Storing lightning inside thunder: Researchers are turning optical data into readable soundwaves, septembre 2017 et enfin Photonic Neuromorphic Computing de Rafatul Faria (35 slides).

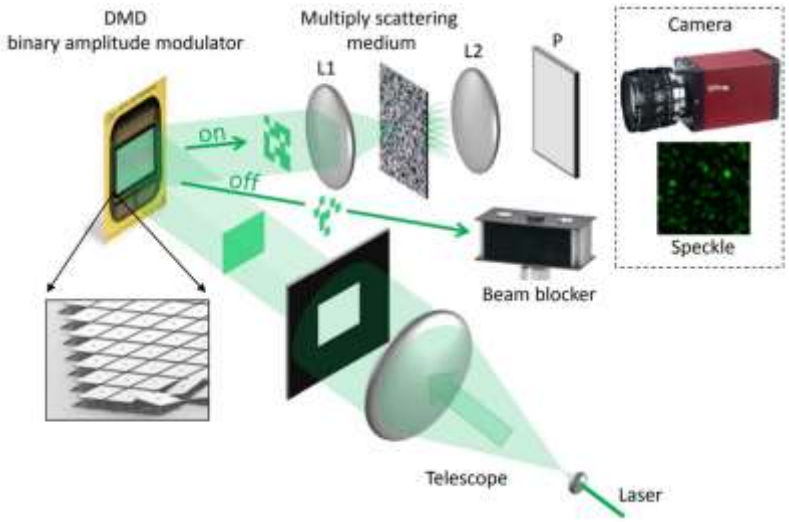
[576] Voir Light-Powered Computers Brighten AI's Future, juin 2017 ainsi que Making AI algorithms crazy fast using chips powered by light, 2019.

[577] Voir aJ/bit Modulators and Photonic Neuromorphic Computing de Volker Sorger, 2017 (15 slides).

[578] Voir Silicon Photonics for Neuromorphic Computing Acceleration of Deep Neural Network training par Folkert Horst, 2018 (18 slides) puis Researchers use light to achieve ultra-low latency for AI systems par Abu Sebastian, janvier 2021 qui fait référence à Parallel convolution processing using an integrated photonic tensor core par Johannes Feldmann et al, octobre 2020 (35 pages).

Lighton.io (2016, France, $3,7M) met au point un coprocesseur optique servant à accélérer d'un facteur 4 à 8 l'entraînement de réseau de neurones sur de gros volumes de données, comme des réseaux convolutifs.

Le procédé est entièrement optique. Un laser émet une lumière qui est agrandie pour éclairer un composant DLP à micro-mirroirs comme dans les projecteurs vidéo et un capteur CMOS qui génère des masques.



La lumière passe ensuite au travers d'un filtre de dispersion puis atterri dans un capteur CMOS monochome. Le système s'appuie sur la génération de jeux de données aléatoires permettant de tester simultanément plusieurs hypothèses de calcul, à des fins d'optimisation[579].

Celui-ci récupère le résultat des interférences générées par l'ensemble et un traitement mathématique permet d'en interpréter le résultat. Le dispositif a été miniaturisé progressivement pendant la mise au point, tenant dans l'équivalent d'un serveur 4U à ce stade. La puissance du système vient en particulier de la résolution du DLP et du capteur CMOS, qui est de plusieurs millions de pixels. Le tout est piloté à partir de bibliothèques Python développées avec TensorFlow. Les applications visées sont en premier lieu la génomique et l'Internet des objets.

Les OPU (Optical Processing Units) de LightOn seraient proposés dans le cloud via les opérateurs français **OVH** et **Scaleway** et directement par LightOn (lien).

LightOn n'est pas le seul sur ce créneau avec quelques autres sociétés plus ou moins concurrentes. Mais comme leur communication est bien moins précise que celle de LightOn, il est très difficile de se faire une opinion[580].

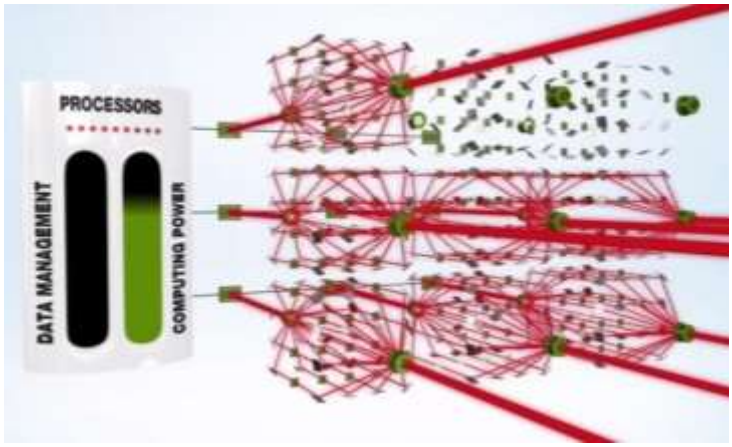
Nous avons donc :

- **Fathom Computing** (2014, USA) qui utilise une architecture "électro-optique" qui est capable d'entraîner des réseaux de neurones à mémoire (LSTM) et convolutifs. Leur Light Processing Unit (LPU) serait capable de lire 90% des tests de la base d'écriture manuscrite MNIST. Dans l'engin, nous avons un laser et des lentilles sans plus de précisions. Le système est adapté à l'algèbre linéaire (ce qui est bizarre lorsque l'on sait que les fonctions d'activation des neurones ne sont généralement pas linéaires) et à la multiplication de matrices qui n'est pas sans rappeller le fonctionnement des ordinateurs quantiques (mais ces derniers utilisent des matrices de nombres complexes). Il leur reste à miniaturiser leur dispositif, ce qui selon eux devrait prendre au moins deux ans. Au passage, ne pas confondre Fathom Computing avec les chipsets Fathom de Movidius, une startup de composant acquise par Intel en 2016 et que nous avons déjà citée précédemment.

[579] Le procédé est décrit dans Random Projections through multiple optical scattering: Approximating kernels at the speed of light, 2015 (6 pages).

[580] Voilà de quoi s'alimenter sur le sujet, avec Progress in neuromorphic photonics par Thomas Ferreira de Lima et al, 2016 (23 pages) et All-optical diffractive neural networks process broadband light par UCLA, décembre 2019.

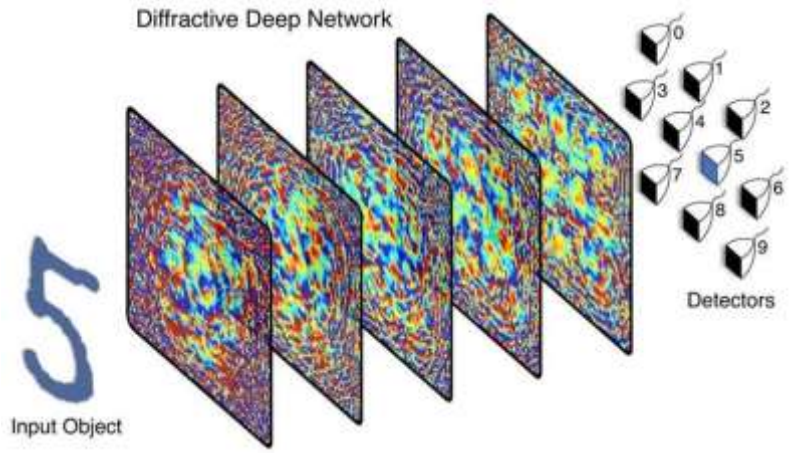
- **Optalysys** (2013, UK, $5,2M)[581] a de son côté mis en oeuvre son premier réseau convolutionnel début 2018 sur une base MNIST avec 60 000 lettres pour l'entraînement et 10 000 pour les tests. Mais avec un taux de réussite de seulement 70%. Leur système permet aussi de réaliser des transformées de Fourier rapides. Ils sont aussi impliqués dans divers projets, l'un en génétique pour faire des recherches de séquences de génomes.



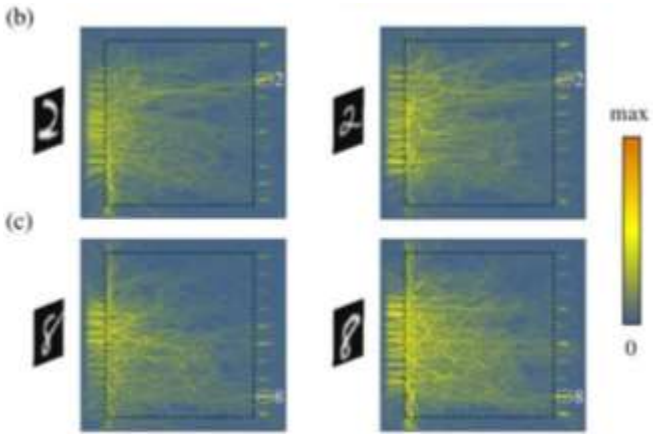
  L'autre pour faire des prévisions météo et un troisième pour la simulation de plasmas et de dynalique des fluides (pour la DARPA).

- **Luminous Computing** (2018, USA, $9M) ambitionne aussi de créer un composant optique ultra-performant. Celui-ci remplacerait 3000 TPUs de Google ! Il exploiterait des lasers de plusieurs couleurs et des guides de lumière[582].

- Un projet de l'Université de Los Angeles (**UCLA**) a l'air d'utiliser un procédé optique qui a un lien de parenté avec celui de Lighton [583], s'appuyant sur un système à diffraction multiple réalisé par impression 3D, créant un réseau de neurones de cinq couches de convolution. Un tel réseau serait statique et non programmable dynamiquement.



- Une autre méthode créé par le **MIT** et l'**Université du Wisconsin** utilise une sorte de réseau photonique « en dur » contenant des nano-éléments diffractant la lumière permettant la reconnaissance de caractères à une vitesse record[584]. Il génère un taux de reconnaissance de 80% sur un jeu de test de 1000 images de lettres.



[581] Voir Optalysys - Revolutionary Optical Processing for HPC, septembre 2017 (23 mn).

[582] Voir The silicon photonics key to building better neural networks, mai 2019, qui référence Design of optical neural networks with component imprecisions 2019 (21 pages) qui a l'air d'être de la même veine que ce que fait Luminous Computing.

[583] Voir This 3D-printed AI construct analyzes by bending light, juillet 2018.

[584] Voir Nanophotonic media for artificial neural inference par Erfan Khoram & Al, 2019 (5 pages). MIT et Université du Wisconsin.

Ordinateurs quantiques

J'ai passé tout l'été 2018 à publier une série d'articles sur le sujet, compilés ensuite dans un ebook publié fin septembre, puis l'été 2019 à le mettre à jour une première fois et l'été 2020 une troisième fois. Nous allons ici en extraire quelques éléments qui concernent l'intelligence artificielle[585].

Les ordinateurs quantiques servent à résoudre des problèmes mathématiques dits exponentiels, dont la complexité et la taille grandit exponentiellement en fonction de leur taille. Ils s'appuient sur des qubits, des unités de gestion d'information manipulant des 0 et des 1, mais en état de superposition, et arrangés dans des registres de plusieurs qubits. Un système à base de n qubits est capable de représenter simultanément 2^n états qui sont des combinaisons de n 0 et 1 sur lesquels diverses opérations peuvent être appliquées simultanément via des portes quantiques et en utilisant le principe de l'intrication pour relier conditionnellement les qubits entre eux.

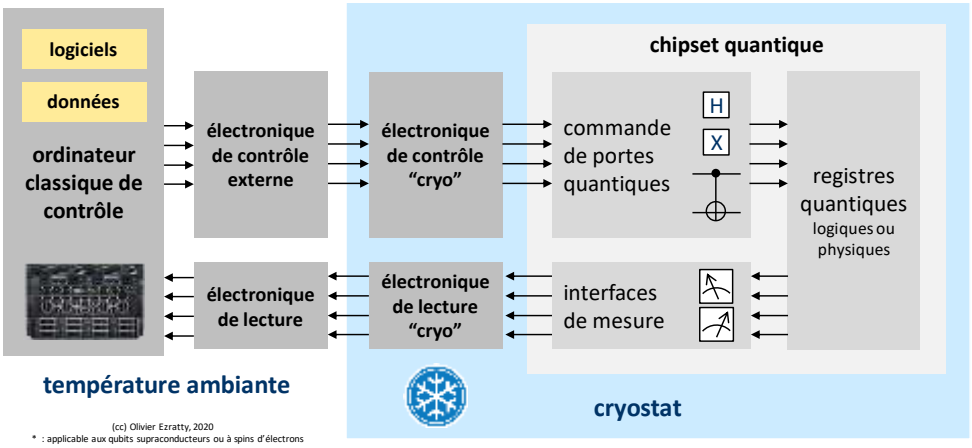
Les qubits sont binaires lors de leur initialisation (en fait, à 0) et au moment où on lit leur valeur après les calculs (0 ou 1). Contrairement à certaines simplifications, les qubits n'ont pas une « capacité de stockage » exponentielle puisque lorsqu'on en lit l'état, on récupère des bits.

Dans de nombreux algorithmes quantiques, au gré de l'augmentation de la complexité du problème, le temps de calcul augmente de manière linéaire ou polynomiale au lieu de grandir exponentiellement comme dans les ordinateurs traditionnels.

Mais les qubits sont complexes à mettre en œuvre : quelle que soit la technologie employée, ils doivent être souvent refroidis à quelques millikelvins au-dessus du zéro absolu avec des systèmes de cryogénie à base d'hélium liquide. Il est surtout difficile de modifier leur état et de le lire. Les qubits sont sujets aux perturbations de l'environnement ambiant qui génèrent des erreurs via le phénomène de la décohérence quantique. Les efforts des physiciens visent à réduire les taux d'erreurs ou à les compenser par des codes de correction d'erreurs. Ceux-ci nécessitent de mettre en œuvre des qubits logiques qui exploitent plusieurs qubits physiques, démultipliant de plusieurs ordres de grandeur le nombre de qubits physiques nécessaires aux calculs.

Un ordinateur quantique est toujours un coprocesseur d'un ordinateur classique, comme décrit dans le schéma *ci-contre*.

Des portes quantiques physiques agissent sur des qubits qui ne bougent pas, sauf pour les qubits à base de photons.



On enchaîne une série de portes quantiques contrôlées par programmation classique sur ces qubits et on lit ensuite le résultat des traitements en évaluant l'état des qubits qui retour un 0 ou un 1.

L'un des premiers algorithmes quantiques apparus qui soit traitable par un ordinateur quantique est celui de **Peter Shor** (AT&T), en 1994. Il permet de factoriser des nombres entiers en nombres premiers avec un temps de calcul qui évolue en fonction du logarithme du nombre plutôt de sa racine carrée dans les calculateurs traditionnels[586].

---

[585] Voir Comprendre l'informatique quantique – édition 2020, septembre 2020 (684 pages).

[586] La version 2017 de cet ouvrage contenait une grosse erreur à ce sujet. J'écrivais au sujet de l'algorithme de Shor « Il permet de factoriser des nombres entiers en nombres premiers avec un temps de calcul qui évolue en fonction du logarithme du nombre plutôt de son exponentielle comme avec les calculateurs traditionnels ». Ce n'était pas « exponentielle » mais « racine carrée ». Nuance !

Il permet de casser les clés publiques utilisées en cryptographie, en particulier avec l'algorithme RSA qui est omniprésent sur Internet. Ce qui le remet sérieusement en question[587] ! Ont suivi divers algorithmes de recherche (1996), d'optimisation (parcours du voyageur de commerce), de simulation de la physique des matériaux et même des mécanismes de la photosynthèse.

Une classe entière d'algorithmes quantiques est dédiée à l'optimisation dans les domaines du machine learning et du deep learning. Le tableau *ci-dessous* positionne clairement les différentes accélérations quantiques associées à divers algorithmes utilisés dans le machine learning et le deep learning[588]. Les accélérations en log(N) sont plus importantes que celles qui sont exprimées en racine carré de N.

Method	Speedup	AA	HHL	Adiabatic	QRAM
Bayesian Inference [107, 108]	$O(\sqrt{N})$	Y	Y	N	N
Online Perceptron [109]	$O(\sqrt{N})$	Y	N	N	optional
Least squares fitting [9]	$O(\log N^{(*)})$	Y	Y	N	Y
Classical BM [20]	$O(\sqrt{N})$	Y/N	optional/N	N/Y	optional
Quantum BM [22, 62]	$O(\log N^{(*)})$	optional/N	N	N/Y	N
Quantum PCA [11]	$O(\log N^{(*)})$	N	Y	N	optional
Quantum SVM [13]	$O(\log N^{(*)})$	N	Y	N	Y
Quantum reinforcement learning [30]	$O(\sqrt{N})$	Y	N	N	N

Il liste de nombreux algorithmes quantiques de bas niveau qui sont très utiles au machine learning et au deep learning[589] : la **PCA** (Principal Components Analysis) pour déterminer les variables clés d'un jeu de données, le **SVM** (support vector machine), une méthode traditionnelle de segmentation dans le machine learning, le **feature mapping** dans le deep learning, pour détecter des formes de manière efficace[590], les **circuits variationnels** (variational circuits), une famille d'algorithmes hybrides qui associent un algorithme quantique et un algorithme traditionnel qui pilote ce dernier[591], des algorithmes quantiques de **réseaux de neurones convolutifs**[592], de taille encore modeste pour l'instant, des algorithmes quantiques de **GAN** (Generative Adversarial Networks) qui génèrent des contenus synthétiques à partir de contenus existants en vérifiant leur plausibilité via un réseau de neurones de reconnaissance[593], un algorithme quantique hybride de **régression non linéaire**, une des méthodes de base de prévisions de valeur quantitative du machine learning[594].

Tous ces algorithmes de réseaux de neurones doivent contourner le fait que les fonctions d'activation des neurones sont généralement non linéaires, comme les sigmoïdes qui sont couramment utilisées alors que les portes quantiques appliquent toutes des transformations linéaires[595]. Il faut aussi trouver des solutions optimales pour alimenter le calculateur quantique avec les données d'entraînement des modèles.

---

[587] Voir Comprendre l'informatique quantique – cryptographie, août 2018.

[588] Le tableau provient de Quantum Machine Learning, mai 2018 (24 pages). Voir aussi cette liste d'algorithmes de machine learning en version quantique dans Quantum Machine Learning What Quantum Computing Means to Data Mining de Peter Wittek, 2014 (178 pages)[588].

[589] Voir Machine learning in the quantum era, Quantonation, novembre 2019.

[590] Voir Supervised learning with quantum enhanced feature spaces, Aram Harrow & Al, 2018 (22 pages) décrit l'usage du quantique pour détecter des formes complexes, bien au-delà de ce que peuvent faire les réseaux de neurones convolutifs ("feature mapping").

[591] Voir Universal Variational Quantum Computation de Jacob Diamonte, 2019 (5 pages).

[592] Voir Quantum Convolutional Neural Networks, par Iris Cong & Al, mai 2019 (12 pages).

[593] C'est bien documenté dans Quantum generative adversarial learning de Seth Lloyd et Christian Weedbrook, 2018 (5 pages) ainsi que dans Quantum generative adversarial learning in a superconducting quantum circuit, 2018 (5 pages).

[594] Voir Nonlinear regression based on a hybrid quantum computer, 2018 (7 pages), issu de chercheurs de plusieurs laboratoires en Chine.

[595] L'astuce est expliquée dans Quantum Neuron: an elementary building block for machine learning on quantum computers, de Yudong Cao, Gian Giacomo Guerreschi et Alan Aspuru-Guzik en 2017 (30 pages).

Ces techniques seront concurrencées par les futurs processeurs neuromorphiques à base de memristors qui permettront de faire converger plus rapidement les réseaux par rétropropagation, comme vu précédemment.

Il existe de nombreuses catégories de processeurs quantiques qui se définissent par leur technologie de qubits.

Les principales sont notamment à base de :

- **Recuit simulé quantique**, ou quantum annealing, chez le canadien **D-Wave** (1999, Canada, $204M) qui est le seul à commercialiser des ordinateurs quantiques à ce jour, avec 5000 qubits dans sa série Advantage lancée en septembre 2020, même si leur efficacité est contestée. Pour les puristes, ce ne sont pas véritablement des ordinateurs quantiques. Mais ils ont l'avantage d'exister et de nombreux algorithmes quantiques de machine learning fonctionnent dessus.
- **Boucles supraconductrices** à effet Josephson, chez **IBM**, **Google**, **Rigetti** (2013, USA, $198,5M) ainsi qu'**Alice&Bob** (France). Le record en date est situé de 65 qubits chez IBM depuis l'été 2020. Google a annoncé en octobre 2019 avoir atteint la suprématie quantique avec le processeur Sycamore de 53 bits et sur un algorithme qui ne sert pas à grand-chose[596].
- **Spin d'élections**, chez **Intel**, ainsi qu'au **CEA-Leti** à Grenoble. Ils présentent l'avantage d'être miniaturisables en tirant parti des capacités de fabrication de la micro-électronique CMOS, et donc, de « scaler », ce qui n'est pas le cas de la plupart des autres technologies de qubits. Ils sont cependant toujours en cours de mise au point. On arrive pour l'instant à intriquer quatre qubits de ce genre.
- **Ions piégés**, comme chez la startup **ionQ** (2016, USA, $82M) issue de l'Université du Maryland et de l'Université Duke en Caroline du Nord ou chez **Honeywell**. Fin 2020, IonQ alignait 32 qubits et Honeywell une dizaine.
- **Atomes froids**, comme chez la startup **Pasqal** (2019, France) dont la technologie permet de dépasser les deux cent qubits, pour l'instant dans un modèle de programmation particulier que l'on appelle la simulation quantique.
- **Qubits topologiques**, chez **Microsoft** avec les fermions de Majorana dont l'existence a été plus ou moins prouvée en laboratoire, et dans les Bell Labs de **Nokia** aux USA. Aucune démonstration d'ordinateur quantique de ce type, même avec un seul qubit n'a pour l'instant été réalisée.

Certains types de qubits sont notamment plus difficiles à stabiliser que d'autres. Les caractéristiques qui déterminent la performance d'un ordinateur quantique sont nombreuses : la première est le taux d'erreur des qubits et des portes qui agissent sur eux, la seconde est le temps de cohérence des qubits, c'est-à-dire, le temps pendant lequel les qubits sont en état de superposition, avec le temps d'exécution des portes quantiques, ces temps et les taux d'erreurs conditionnent le nombre de portes quantiques qui pourront être enchaînées dans un algorithme.

Lorsque les ordinateurs quantiques avec 50 à 100 qubits verront le jour, il est probable que l'on assistera à un bon développement de leurs domaines d'applications. À quelle échéance ? L'incertitude est très grande à ce sujet avec des prévisions qui s'étalent entre 5 ans et… jamais ! L'échéance de la singularité par l'AGI est plus déterministe de ce point de vue-là, même si les prévisions associées (2045 ou autre) ne veulent rien dire.

À plus long terme, le calcul quantique pourra concurrencer l'IA, en ce sens qu'il permettra de simuler des systèmes physiques à partir de leurs lois de fonctionnement quantiques. Cela pourra remplacer des mécanismes de prévision à base de deep learning qui génèrent des extrapolations à partir d'observations existantes.

---

[596] Voir Interpréter la suprématie quantique de Google, Olivier Ezratty, septembre/octobre 2019.

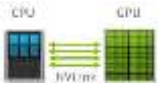
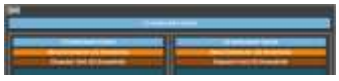
Mémoire

Les GPU et les processeurs neuromorphiques sont d'autant plus performants dans les phases d'entraînement qu'ils accèdent rapidement aux données en mémoire, et notamment aux paramètres des réseaux de neurones qui peuvent être des dizaines de millions voire des milliards de variables à ajuster très fréquemment.

Pour cela, les technologies matérielles rapprochent de plus en plus les unités de traitement de mémoires de plus en plus rapides[597]. Un serveur peut avoir jusqu'à une demi-douzaine de niveaux de mémoire qui optimiseront la performance de l'ensemble. Sachant que plus la mémoire est rapide, plus elle est coûteuse et plus sa taille est limitée. Nous avons donc une hiérarchie de mémoires dont la vitesse augmente inversement proportionnellement à leur taille.

Le cas le plus extrême est celui d'une mémoire non volatile directement intégrée au sein des neurones comme nous l'avons creusé dans la partie sur les memristors. Pour certains scientifiques, c'est la seule voie de salut pour continuer à faire progresser les architectures matérielles de l'IA[598].

Nous allons faire le tour des principaux types et niveaux de mémoire qui équipent aujourd'hui les serveurs. Nous nous focalisons sur les serveurs qui réalisent l'entraînement de réseaux de neurones car ils sont les plus sollicités en calcul. L'inférence d'un réseau de neurones déjà entraîné est très simple et peu consommateur de calcul par rapport à son entraînement.

		vitesse max	capacité
SSD M.2 PCIe stockage		3 Go/s	>1 To
DDR4 mémoire externe CPU		10 Go/s	>16 Go
Infiniband comm interserveur	 	25 Gos/s	*bus de données*
GDDR6X mémoire externe GPU		1 To/s	2 Go – 12 Go
NVLink 2.0 comm inter-GPU/CPU		900 Go/s	*bus de données*
HBM2e / HMC mémoire externe GPU		1,8 To/s	16 Go
GPU cache & registres mémoire interne GPU		> 16 To / s	6 Mo (L1)

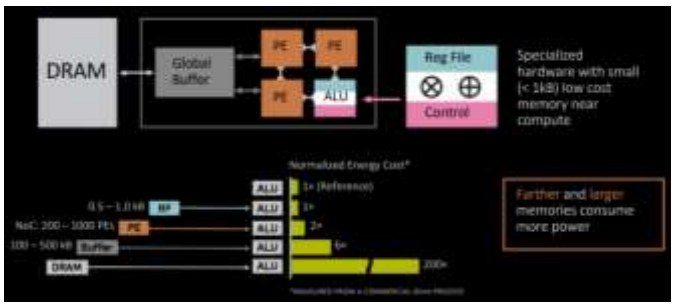
L'enjeu dans l'embarqué est de réaliser ces inférences en consommant le moins d'énergie possible. Cette énergie est surtout consommée dans les accès mémoire, bien plus que dans le calcul, et plus la mémoire est éloignée du traitement, plus son coût d'accès augmente[599].

---

[597] Cela concerne aussi les logiciels eux-mêmes qui peuvent optimiser le chargement des données en mémoire pendant les phases d'entraînement. Voir à ce sujet Powerful AI Can Now Be Trained on a Single Computer par Edd Gent, juillet 2020, qui fait référence à Sample Factory: Egocentric 3D Control from Pixels at 100000 FPS with Asynchronous Reinforcement Learning par Aleksei Petrenko, 2020 (18 pages).

[598] Voir AI Architectures Must Change de Brian Bailey, août 2018 et 3D Neuromorphic Architectures, de Katherine Derbyshire, décembre 2017.

[599] Voir cette présentation de Vivienne Sze, avril 2019 (19 slides) d'où sont extraites les comparaisons de coût énergétique d'accès à la mémoire et de traitements. Voir aussi DNN Dataflow Choice Is Overrated par Mark Horowitz & Al, septembre 2018 (13 pages) qui décrit diverses techniques d'optimisation d'accès à la mémoire pour l'accélération de réseaux de neurones.



Mémoire cache

Au sein des processeurs se trouve de la mémoire cache volatile qui est utilisée directement par les unités de traitement. En technologie CMOS classique, sa vitesse d'accès est ce qui se fait de plus rapide, et dépasse les To/s (téra-octets par seconde).

Un processeur courant comprend précisément deux à trois niveaux de cache et des registres mémoire. Plus on se rapproche des unités de traitement, plus l'accès à cette mémoire cache est rapide, mais plus elle est limitée en capacité, de l'ordre de quelques dizaines de Ko, soit juste de quoi alimenter les registres de calculs utilisés dans les processeurs et de quoi en lire les résultats.

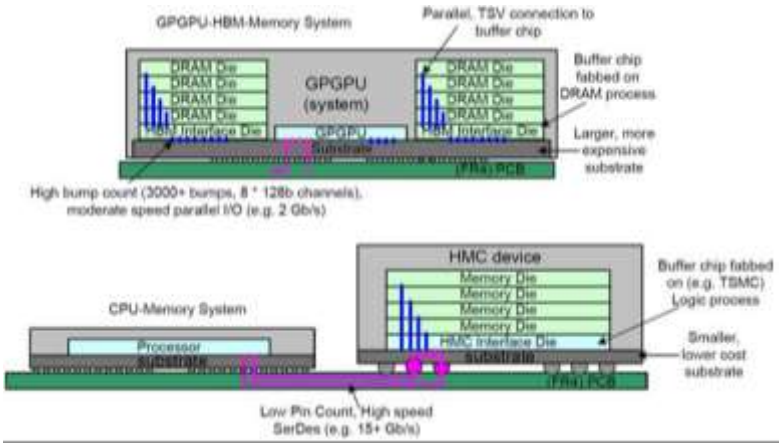
La mémoire intégrée dans les processeurs est limitée à quelques Mo. Pourquoi donc ? Parce qu'elle est chère à fabriquer et qu'il n'y a pas beaucoup de place sur les puces des processeurs.

HBM2 et HMC

Autour des GPU et de certains TPU (Tensor Processing Units) se trouve maintenant souvent une mémoire complémentaire très rapide utilisant l'un des deux grands standards du marché **HBM2** (High Bandwidth Memory) ou **HMC** (Hyper Memory Cube).

Le premier standard est promu par AMD et le Coréen SK Hynix et le second par l'Américain Micron avec le support de Samsung. Cette mémoire qui atteint aujourd'hui 32 Go est située dans des circuits intégrés empilés par paquets de 4 ou 8 et reliés entre eux et avec le GPU ou le CPU par des micro-conducteurs métalliques.

Cette proximité permet d'avoir une vitesse rapide de transfert d'information entre la mémoire et le processeur. L'intégration avec le GPU est plus étroite pour le HBM2 car la mémoire et le GPU sont installés sur un substrat commun tandis que pour le HMC, la mémoire est placée sur la carte mère au même titre que le CPU[600]. Ces mémoires permettent d'atteindre des débits allant jusqu'à 900 Go/s dans le Nvidia Volta V100 lancé en 2017.



HBM2 est utilisé dans les GPU Nvidia V100 et HMC l'est dans les processeurs serveurs Intel Xeon Phi ainsi que dans les FPGA Intel Stratix 10MX utilisés notamment par Microsoft dans ses processeurs neuromorphiques Brainwave.

[600] Source du schéma qui suit : A Talk on Memory Buffers, Inphi.

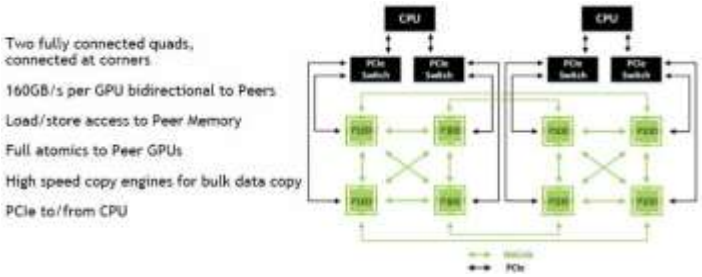
En août 2019, **SK Hynix** annonçait des puces HBM2e permettant d'atteindre un débit de 1,8 To/s, doublant la performance obtenue auparavant[601]. Il reste à les intégrer de bout en bout dans des processeurs de nouvelle génération. Peut-être pour le futur successeur des GPU V100 de Nvidia qui pourraient voir le jour en 2020 ou 2021.

NVLink

La technologie **NVLink** de Nvidia permet de relier les GPU entre eux ou les GPU avec les CPU à une vitesse atteignant 300 Go/s par composant. La version 2.0 de NVLink utilisé des chipsets dédiés, les NVSwitch, donnant un débit total de 900 Go/s.

Cette connexion permet de répartir optimalement les traitements parallèles sur plusieurs processeurs.

En effet, les grands modèles de réseaux de neurones doivent être répartis sur plusieurs GPU et plusieurs serveurs. Ils peuvent être jusqu'à des milliers mis en batterie dans le cloud !



Infiniband

Infiniband est une technologie permettant de relier les serveurs entre eux avec des débits compris entre 100 et 200 Gbits/s. La connexion se fait via un câble différent du RJ45 des réseaux Ethernet.

Les composants supportant Infiniband sont commercialisés par **Mellanox Technologies** (1999, Israël, $89M) et par Intel. Infiniband est notamment complété par le standard RoCE (RDMA over Converged Ethernet) qui permet de permettre l'accès à la mémoire d'un serveur par un autre serveur. Infiniband est concurrencé par Fibre Channel, une autre technologie de liaison entre serveurs qui peut atteindre 128 Gbits/s et sert surtout à l'optimisation de l'accès au stockage. Toutes ces technologies sont utilisées dans les data centers et les super-ordinateurs (HPC).

En mars 2019, **Nvidia** annonçait l'acquisition de l'Israélien **Mellanox**, un spécialiste des communications haute vitesse dans les datacenters et expert dans la technologie Infiniband qui permet de relier des serveurs à distance via fibre optique**.** Et pour $6,9B. Grâce à cela, Nvidia contrôle une bonne part de la chaîne de valeur des communications inter-CPU/GPU et serveurs dans les datacenters.

GDDR

La mémoire **GDDR** est utilisée dans les cartes graphiques et est plus rapide que la mémoire DDR4 qui est exploitée dans les micro-ordinateurs. Elle atteint une bande passante de 48 Go/s dans sa version 5.0 et 64 Go/s dans sa version 6.0. Les chipsets en version 6.0 sont apparus en 2018. Les puces de mémoire GDDR6 proviennent des habituels fournisseurs de mémoire que sont **Samsung**, **SK-Hynix** et **Micron**[602].

Nvidia a intégré de la mémoire GDDR 6 dans ses cartes graphiques GeForce RTX 2080 Ti, RTX 2080 et RTX 2070 en architecture « Turing », lancées en août 2018.

---

[601] Voir SK Hynix Announces 3.6 Gbps HBM2E Memory For 2020: 1.8 TB/sec For Next-Gen Accelerators, août 2019. Voir aussi HBM2E Memory: A Perfect Fit For AI/ML Training par Franck Ferro, mars 2020.

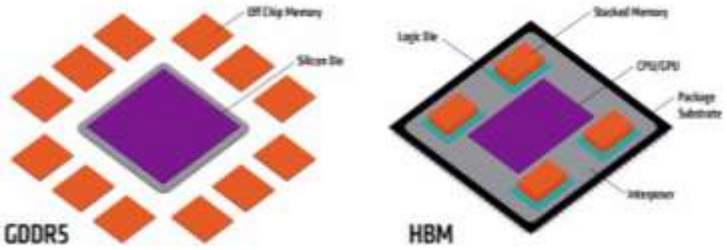
[602] Voir GDDR6 Pushes The Memory Envelope For AI And ADAS par Frank Ferro, novembre 2019.

Le nouveau standard GDDR6X lancé en 2020 par Micron et Nvidia équipe les cartes graphiques GeForce RTX 3080 et 3090 et atteint une bande passante de 1 To/s.

Celles-ci visent surtout le marché du jeu vidéo et se distingent par leurs capacité de « ray tracing » temps-réel pour la génération d'images 3D photoréalistes, mais comprennent tout de même des tenseurs pour l'exécution d'applications de deep learning, notamment pour l'upscaling d'images en 4K.

Ces chipsets peuvent exécuter 500 trillions d'opérations par seconde dans ces tensors, un niveau difficile à comparer avec d'autres processeurs[603].

La mémoire GDDR est remplacée par de la mémoire HMC ou HBM2 depuis quelques temps dans les GPU haut de gamme dédiés aux serveurs. Ces nombreuses avancées montrent que les fabricants de ces composants ont encore du mou sous la pédale. Peu d'obstacles les empêchent ainsi, à moyen terme, d'intégrer de plus grandes capacités de mémoire rapide dans les processeurs eux-mêmes.



C'est une question de maîtrise de la fabrication de circuits intégrés de grande taille et avec des dizaines de milliards de transistors.

In memory processing

Citons pour terminer sur la mémoire un cas original avec la startup fabless **Upmem** (2015, France, 3M€), basée à Grenoble, la capitale française des nanotechnologies, qui intègre des unités de traitement DPU (DRAM Processing Units) directement à l'intérieur de mémoires DRAM, permettant une accélération de certains traitements applicables notamment au data mining. L'idée, issue du CEA-Leti, consiste à intégrer dans des chipsets de mémoire des unités de traitement RISC (jeu d'instruction simple) 32 bits, le dimensionnement pouvant être de 256 unités de traitement dans des chipsets de 16 Go de RAM. Bref, au lieu de mettre de la mémoire rapide dans des processeurs, ils mettent des unités de traitement dans la mémoire rapide ! Ces DRAM actives sont des coprocesseurs de traitement de CPU traditionnels. Reste à les fabriquer en volume et à les faire intégrer dans des serveurs par leurs constructeurs[604] !

Dans le même ordre d'idée, **WITINMEM** (2017, Chine, $15M) développe des chipsets d'inférence qui rapprochent le calcul de mémoire flash. Et **Samsung** annonçait début 2021 intégrer des modules de calcul de machine learning dans des puces mémoires HBM2 dénommées HBM-PIM.

GenZ, CXL et CCIX

Divers consortiums ont été lancés ces dernières années pour structurer et accélérer les échanges entre les CPU, GPU et autres accélérateurs et la mémoire.

---

[603] Voir les détails sur les tenseurs des GPU Turing : The NVIDIA Turing GPU Architecture Deep Dive: Prelude to GeForce RTX, septembre 2018. Ils supportent 114 TFLOPS en FP16 (flottant 16 bits) et 455 TOPS en entiers 4 bits. Le processeur comprend 18,9 milliards de transistors, un peu moins que les 21 milliards du V100 lancé en 2017 et destiné aux serveurs.

[604] Voir la description détaillée de la technologie UPMEM dans Hot Chips 31 Analysis : In-Memory Processing by UPMEM, août 2019. Et plus généralement, une comparaison des architectures consistant à rapprocher la mémoire du calcul, dans Near-Memory Computing: Past, Present, and Future par Gagandeep Singh et al, août 2019 (16 pages).

GenZ est une initiative lancée par HPE pour redéfinir l'interconnexion mémoire dans les supercalculateurs issu du projet The Machine lancé en 2014 qui devait révolutionner le calcul haute performance avec des memristors et des communications optiques. C'est un protocole unique prêt pour des interconnexions optiques de bout en bout permettant de réduire la latence et la consommation d'énergie, qui ne dépend pas de la distance. Le protocole peut servir à créer des architectures de machine learning massivement parallèles et distribuées gérant un grand nombre de paramètres de réseaux de neurones.

CXL (Compute Express Link) est un autre consortium qui a spécifié un modèle d'accès partagé à la mémoire basé sur le bus PCIe pour gérer des entrées/sorties, la mémoire et le cache entre divers accélérateurs. Là aussi, cela peut servir au machine learning. Les membres fondateurs comprennent Alibaba, Cisco, Dell EMC, Facebook, Google, HPE, Huawei, Intel et Microsoft.

CCIX (Cache Coherent Interconnect for accelerators) est un consortium concurrent de CXL, lancé en 2017 par Xilinx et arm, qui a défini une spécification d'interconnexion avec des accélérateurs et de gestion de cache sous forme d'une extension du bus PCIe[605]. Dès qu'un standard est créé, un standard concurrent est lancé à côté et c'est la zizanie. Même s'il existe bien entendu des différences techniques entre les deux.

Stockage

Le stockage d'informations se fait de plus en plus dans des SSD, les disques de stockage sans plateau mobile et à base de mémoire flash NAND et V-NAND. Ce sont des circuits intégrés à plusieurs couches, empilant jusqu'à une centaine de couches de transistors. Les SSD grand public atteignent aujourd'hui une capacité de 4 To avec une vitesse d'accès de 3 Go/s dans le meilleur des cas (sur certains laptops Apple). Il existe des SSD de bien plus grande capacité qui sont destinés aux serveurs de data-centers.

Si la loi de Moore a tendance à se calmer du côté des processeurs CMOS, elle continue de s'appliquer au stockage. Elle s'est appliquée de manière plutôt stable aux disques durs jusqu'à présent. Le premier disque de 1 To (Hitachi en 3,5 pouces) est apparu en 2009 et on en est maintenant à 18 To en 2020 et bientôt 20 To chez Seagate. A peu près 2^4, la loi de Moore est donc presque sauve avec un doublement tous les deux ans et demi ! Le progrès s'est ensuite déplacé vers les disques SSD à mémoires NAND dont la capacité augmente régulièrement tout comme sa vitesse d'accès, le tout avec une baisse régulière des prix.

Les perspectives de croissance sont ici plus optimistes qu'avec les processeurs CMOS. Les records côté SSD sont de 30,72 To chez Samsung en 2,5 pouces en février 2018 et de 100 To chez **Nimbus Data**, en 3,5 pouces en mars 2018, sans évolution depuis d'ailleurs. Le SSD de 100 To est commercialisé à $40K.

L'augmentation de la densité des mémoires NAND profite des architectures en trois dimensions qui sont maintenant courantes, comme avec les V-NAND de **Samsung** qui sont utilisées dans leurs SSD pour laptops, desktops et serveurs. Nous avons aussi **Toshiba** (*ci-dessous*) avec sa technologie BiCS. Les puces de mémoire 3D comprennent avec plusieurs couches empilées de transistors (*ci-dessus* à droite), ou de transistors montés en colonnes.

Le niveau d'intégration le plus bas des transistors approche celui des CPU les plus denses : il descend jusqu'à 10 nm. On sait empiler aujourd'hui jusqu'à 136 couches de transistors, notamment dans la 6e génération de V-NAND de **Samsung**.

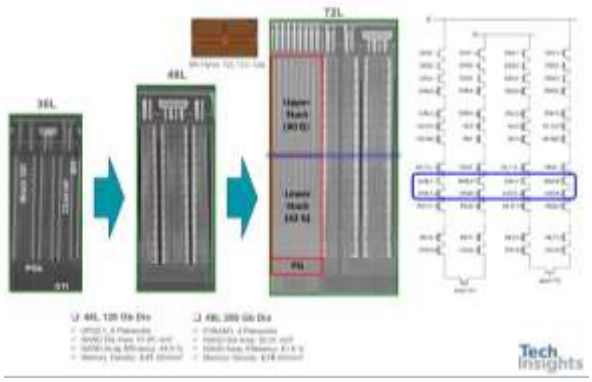
La technologie 3D XPoint d'**Intel** et **Micron** qui combine le stockage longue durée et une vitesse d'accès équivalente à celle la mémoire RAM associée aux processeurs est aussi prometteuse même si elle a connut un double retard à l'allumage : côté disponibilité comme côté performance.

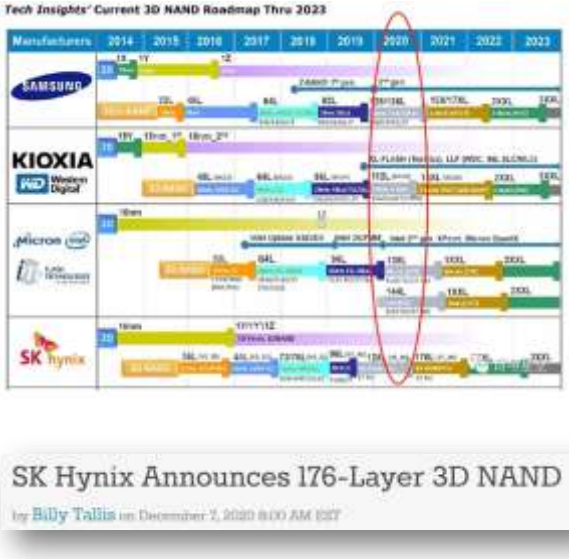
---

[605] Voir CCIX Enables Machine Learning par Brian Bailey, 2017.

On la trouve sur le marché à des tarifs nettement plus élevés que les SSD.

V-NAND:
- >100 layers in production
- limited write cycles
- fast R/W, role of controller





Pourquoi cette intégration verticale est-elle possible pour la mémoire et pas encore pour les processeurs (GPU, CPU) ? C'est visiblement lié à la résistance à la montée en température. Dans un processeur, une bonne part des transistors fonctionne en même temps alors que l'accès à la mémoire est séquentiel et donc n'active pas simultanément les transistors. Un processeur chauffe donc plus qu'une mémoire. Si on empilait plusieurs couches de transistors dans un processeur, il se mettrait à chauffer bien trop et s'endommagerait. Par contre, on sait assembler des circuits les uns sur les autres pour répondre aux besoins d'applications spécifiques.

Ce modèle de mémoire en 3D est également appliqué à la RAM, notamment par l'américain **Micron** avec sa technologie Hyper Memory Cube (HMC, déjà évoquée) ainsi qu'avec la mémoire HBM2. Pour les supercalculateurs, une tâche ardue est à accomplir : accélérer la vitesse de transfert des données du stockage vers les processeurs au gré de l'augmentation de la performance de ces derniers. Cela va aller jusqu'à intégrer de la connectique à plusieurs centaines de Gbits/s dans les processeurs. Mais la mémoire ne suit pas forcément.

Un SSD connecté en PCI et avec un connecteur M.2 est capable de lire les données à la vitesse vertigineuse de 3,2 Go/s, soit un dixième de ce qui est recherché dans les calculateurs à haute performance (HPC). Avec 3D XPoint, l'accès aux données est 1000 fois plus rapide qu'avec les SSD actuels, modulo l'interface utilisée. La technologie aura probablement un impact important pour les systèmes d'IA temps réel comme IBM Watson.

Rappelons-nous que pour Jeopardy, l'ensemble de la base de connaissance était chargée en mémoire RAM pour permettre un traitement rapide des questions [606]!

Cette augmentation de la rapidité d'accès à la mémoire, qu'elle soit vive ou de longue durée, est indispensable pour suivre les évolutions à venir de la puissance des processeurs avec l'une des techniques que nous avons examinées juste avant. Et d'autres obstacles sont à surmonter [607]!

[606] IBM Watson avait chargé en mémoire tout Wikipedia et les questions lui étaient soumises par écrit et pas oralement. Bref, Watson et les joueurs homo-sapiens ne jouaient vraiment pas à armes égales !

[607] Source du schéma : Why we need Exascale and why we won't get there by 2020, 2014 (56 pages).

Cela rappelle le besoin d'équilibrer les architectures de systèmes de plus en plus performants. Mais, comme c'est de la chimie, elle sera probablement plus lente que les changements de phase ou de magnétisme qui ont court dans les systèmes de stockage numérique actuels. La loi de Moore patientera donc quelques décennies de ce côté là, tout du moins pour ses applications dans le cadre de l'IA.

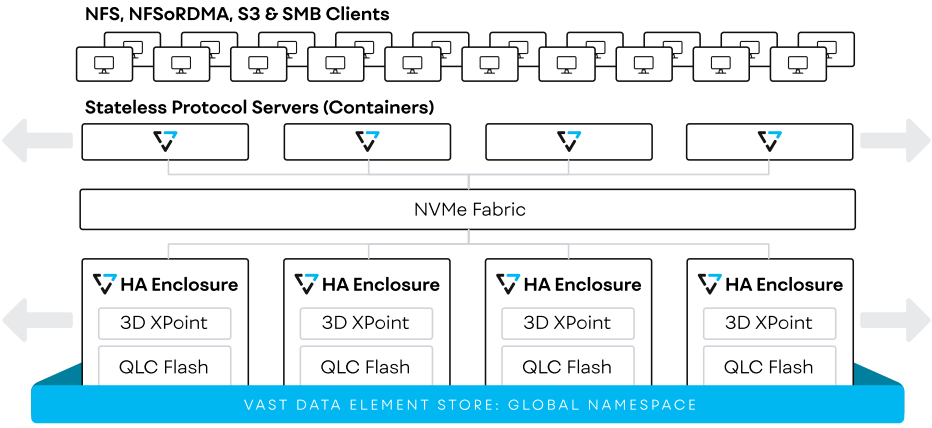
Old Constraints	New Constraints
• **Peak clock frequency** *as primary limiter for performance improvement*	• **Power** *is primary design constraint for future HPC system design*
• **Cost:** *FLOPs* are biggest cost for system: *optimize for compute*	• **Cost:** *Data movement dominates: optimize to minimize data movement*
• **Concurrency**: *Modest growth of parallelism by adding nodes*	• **Concurrency:** *Exponential growth of parallelism within chips*
• **Memory scaling**: *maintain byte per flop capacity and bandwidth*	• **Memory Scaling:** *Compute growing 2x faster than capacity or bandwidth*
• **Locality**: *MPI+X model (uniform costs within node & between nodes)*	• **Locality**: *must reason about data locality and possibly topology*
• **Uniformity**: *Assume uniform system performance*	• **Heterogeneity**: *Architectural and performance non-uniformity increase*
• **Reliability**: *It's the hardware's problem*	• **Reliability**: *Cannot count on hardware protection alone*

Du côté des data centers, l'augmentation de la densité du stockage va bon train avec le SSD. Le top de l'état de l'art ? Des baies de stockage ou des serveurs utilisant des barres de SSD exploitant le connecteur EDSFF (Enterprise and Datacenter SSD Form Factor)[608]. Une baie 1U peut contenir 32 systèmes de stockage NVMe de 32 To, comme ceux d'Intel, et totalisant donc 1 Po. Ils sont plus efficaces pour la densité et le refroidissement. De telles baies sont proposées par différents constructeurs tels que Samsung et SuperMicro. Le SuperMicro 1U Super Server 1029P-NR32R *ci-dessous* comprend 6 processeurs Intel Xeon avec 6 To de RAM et 32 Intel Ruler à connecteurs EDSFF échangables à chaud.

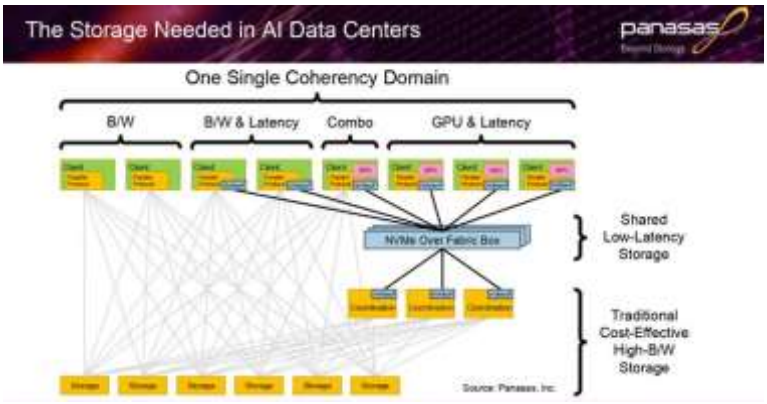
Des solutions d'optimisation de la gestion du stockage sont proposées comme celles de **Vast Data** (2016, USA, $180M) qui combine du stockage SSD classique et du stockage 3D XPoint d'Intel qui est d'accès plus rapide et est utilisé comme buffer d'écriture de données[609]. Leur architecture DASE (Disaggregated Shared Everything) utilise des serveurs qui ont accès aux ressources de stockage directement via le bus PCI.

[608] L'illustration sur les connecteurs EDSFF est extraite de slides de Lenovo dans G2M Research Webinar onComputational Storage and EDSFF for Flash Devices, novembre 2018 (43 slides).

[609] Voir Universal Storage Explained, Vast, 2020.

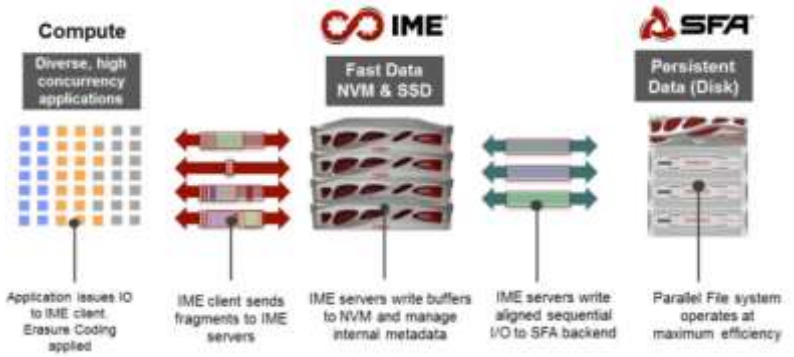


Pour ce qui est d'augmenter la performance du stockage, une autre voie est explorée, notamment par **NGD Systems** (2013, USA, $22,7) : le « computational storage ». Il consiste à intégrer des capacités de calcul dans les unités de stockage via des contrôleurs ASIC dédiés[610]. Cela requiert évidemement une architecture logicielle adaptée.



Une autre startup, **BlueShift Memory** (UK), issue de Cambridge et vue à Hello Tomorrow à Paris en mars 2019, développe des modules d'accès mémoire censés améliorer d'un facteur 100 à 1000 certains calculs, notamment dans le deep learning. Leur prototype est pour l'instant bâti sur un FPGA Xilinx et de la mémoire DRAM dont ils améliorent l'indexation. Ils n'en disent pas plus.

Citons aussi la solution IME (Infinite Memory Engines) de **DataDirect Networks** (1998, USA, $9,9M) qui joue le rôle d'un cache en mémoire flash (SSD) pour du stockage en disque dur classique[611]. Ils complètent cela avec DDN A³I, une solution de stockage adaptée aux applications de l'IA, codéveloppée avec Nvidia et HPE. Ce sont des concurrents de NetApp.



Enfin, citons une voie de stockage d'information assez éloignée des besoins de l'IA malgré les apparences, celle de l'**ADN**. Elle est explorée par des chercheurs d'Universités et chez Microsoft Research. Ce dernier travaille en partenariat avec la startup américaine **Twist Bioscience** (2013, USA, $253M, IPO en 2018) créé par la française Emily Leproust. Plusieurs startups se sont même lancées sur ce créneau (**Kilobaser**, **Helixworks**, **Catalog Infinite Data Archives**, **Evonetix**).

---

[610] Le schéma qui provient de Storage underpins interplay between HPC and AI par Michael Feldman, mai 2019. Voir aussi Building In-Storage Compute - Computational Storage Solutions par NGD Systems, décembre 2018 (15 slides) et Deployment of In-Storage Compute, Scott Shadley, NGD Systems, septembre 2018 (29 slides).

[611] Voir Infinite Memory Engine - A Scale-Out Storage Cache for the Flash Era (10 pages).

Les premières expériences menées depuis quelques années sont prometteuses[612]. La densité d'un tel stockage serait énorme. Son avantage est sa durabilité, estimée à des dizaines de milliers d'années, voire plus selon les techniques de préservation. Reste à trouver le moyen d'écrire et de lire dans de l'ADN à une vitesse raisonnable. Aujourd'hui, on sait imprimer des bases d'ADN à une vitesse incommensurablement lente par rapport aux besoins des ordinateurs. Cela se chiffre en centaines de bases par heure au grand maximum. Cette vitesse s'accélérera sans doute dans les années à venir mais on est très loin des besoins opérationnels de l'IA[613].

Big data

La notion de big data est apparue progressivement entre le milieu des années 1980 et le début des années 2000. L'expression a germé sporadiquement entre 1984 et 1987 mais au premier degré : « beaucoup de données », en relation à une application spécifique comme la création de listes d'envoi de courriers publicitaires. Le big data comme domaine informatique est attribuable en 1996-1998 à **John Mashey**, à l'époque ancien chief scientist de SGI à la retraite. Il apparaît aussi dans la littérature scientifique, notamment autour de l'analyse statistique des données[614].

Concepts

La notion de big data ne recouvre pas simplement la gestion de gros volumes de données. Le concept est notamment affiné par un analyse de META Group, **Doug Laney**, dans une note publiée en février 2001 définissant trois paramètres clés définissant le ou la « big data » dans le « framework 3V » : le (grand) **volume** des données, leur **vélocité** (vitesse de production et de mise à jour) et la **variété** des données (structurées ou pas)[615]. Ces paramètres correspondent à l'émergence de sites web de commerce en ligne qui doivent gérer des données en volume croissant, et avec une forte réactivité.

La notion de big data est cependant est à géométrie variable[616]. Le framework 3V a été étendu dans différentes directions pour devenir successivement 4V, 5V, 7V et 10V, en ajoutant des caractéristiques telles que variabilité, véracité, validité, vulnérabilité, valeur, volatilité, visualisation et valeur des données. Le modèle 10V créé par Kirk Born en 2014 a été modifié depuis (schéma *ci-dessous*)[617]. Il existe même un modèle 17V qui ajoute la viscosité, la viralité, la « venue » (origine des données), le vocabulaire, la vagueté, la verbosité, le volontarisme et la versatilité, en retranchant la vulnérabilité[618]. Bref, hors du V, point de salut ! Le V est le 42 du big data !

---

[612] Sachant néanmoins qu'elles ont démarré en 1994 avec les travaux de Leonard M. Adleman aux USA, documentés dans Computing with DNA paru dans Scientific American en 1998. A cette époque, Adleman voulait créer un ordinateur à base d'ADN. Mais sa conclusion était que l'ADN était surtout un moyen intéressant de stockage de gros volumes d'information. J'ai remarqué au passage dans l'article que le coût de la génération de molécules d'ADN était déjà relativement bas à cette époque : $1,25 la paire de bases d'ADN. Il démarre en 2016 à $0,2, soit seulement 6 fois moins. En plus de 20 ans ! Encore un exemple où la loi de Moore ne s'est pas du tout appliquée. Pour l'instant !

[613] Voir Archiver les mégadonnées au-delà de 2040 : la piste de l'ADN, Académie des Technologies, octobre 2020.

[614] Voir la thèse Phénomène Big Data en entreprise : processus projet, génération de valeur et médiation homme-données, Anna Nesvijevskaia, octobre 2020 (420 pages), pages 24 et suivantes.

[615] Voir "A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline", Second Version par Francis Diebold, 2012 (8 pages), Deja VVVu: Gartner's Original "Volume-Velocity-Variety" Definition of Big Data par Doug Laney, 2012.

[616] Comme le souligne Christophe Benavent dans Big data : no best way, décembre 2014 (11 pages).

[617] Voir Top 10 List – The V's of Big Data, Kirk Borne, avril 2014, Big Data: The V's of the Game Changer Paradigm, par Ripon Patgiri, et Arif Ahmed, décembre 2016 (9 pages) et Voir Application of Big Data Analytics in Customization of E-mass Service: Main Possibilities and Obstacles par Gedas Baranauskas, 2019 (11 pages).

[618] Voir The 17 V's Of Big Data par Arockia Panimalar et al, 2017 (5 pages).

Le phénomène du big data accompagne l'augmentation du volume des données gérées par les entreprises, en particulier pour celles qui ont de gros volumes d'affaires dans des ventes grand public, sur Internet ou en relation avec des objets connectés.

Les données exploitées sont aussi bien directement issues de l'activité humaines (emails, documents, photos, activités dans les réseaux sociaux, déplacements, achats, consommation de médias) que de machines (données issues de capteurs divers, d'instruments d'observations scientifiques, logs de serveurs, séquençage de génomes, etc). On y trouve des données nominales (qui ne peuvent pas être triées comme des pays ou des métiers), ordinales (qui sont triables mais sans unité de mesure), des intervalles et des ratios (mesure, continue, distance calculable, triable). Les volumes se chiffrent en pétaoctets et au-delà.

Les volumes de données les plus importants sont gérés par les grands acteurs de l'Internet avec des pétaoctets accumulés chaque jour et des datacenters qui comprennent maintenant des exaoctets de données. Le monde de la recherche scientifique est aussi génération de très gros volumes de données, comme dans la génomique ou la physique des particules. Le LHC CERN génère de son côté 15 Po par an. Tout cela a amené à considérer que la donnée était le nouvel or noir de l'ère numérique[619].

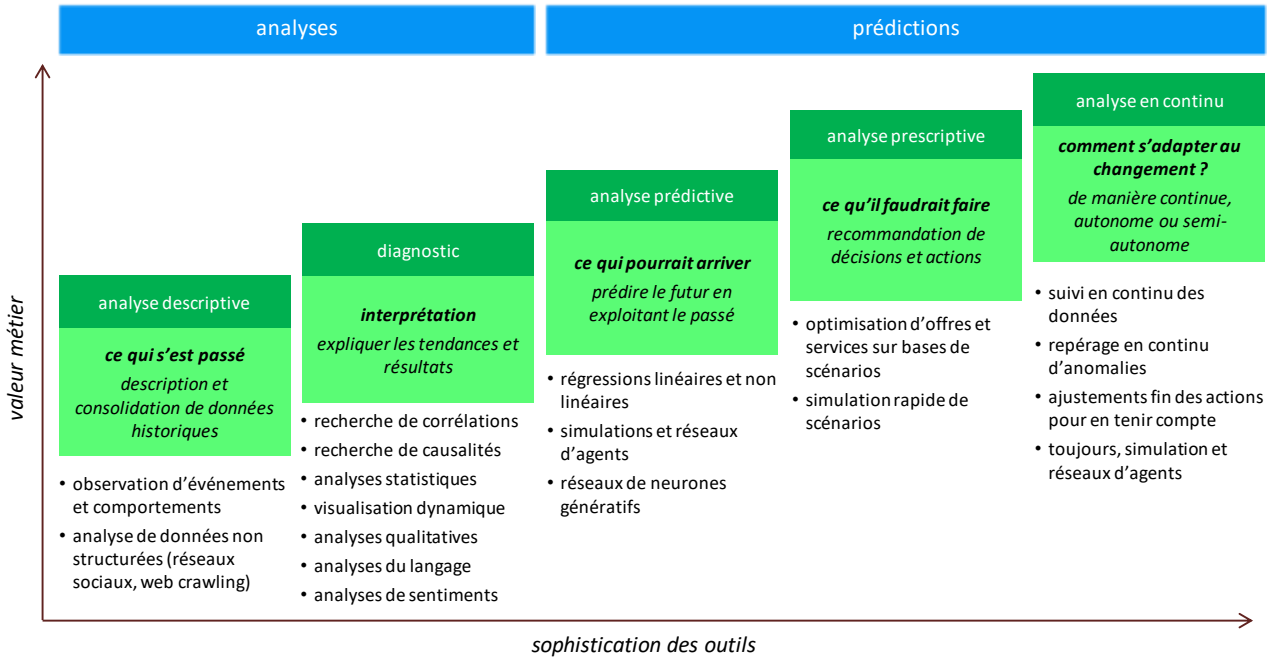
Mais à quoi bon accumuler toutes ces données ? La promesse business est d'accélérer les processus métiers et en particulier, d'identifier des corrélations entre données, de comprendre les lois statistiques sous-jacentes du business et de prédire dans une certaine mesure l'avenir. C'est là que l'on se rapproche inexorablement des outils du machine learning.

Le champ du big data comprend les **big data analytics** (BDA) et le **data mining** qui en sont les principaux outils d'exploitation des données. Le data mining ou l'exploration des données permet l'analyse automatique ou semi-automatique de grandes quantités de données pour en extraire des règles et tendances. Le machine learning en fait partie et constitue une brochette d'outils d'exploitation des gros volumes de données du big data qui fait essentiellement partie des outils des Big Data Analytics.

---

[619] « *Data is the new oil* » selon le mathématicien et entrepreneur britannique Clive Humby, en 2006. C'est évidemment une vision très simpliste des choses car la donnée ne tombe pas du ciel et résulte de la fourniture de services. Les GAFA n'auraient pas les données qu'ils accumulent dans les services qui ont permis de les collecter. Ceci dit, il ajoutait « *It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value* ».

Les outils de big data analytics comprennent l'analyse de textes (text mining, pour classifier des documents, suivre l'évolution de l'opinion, etc), le ciblage marketing, les analyses de risques dans la finance et les assurances, la détection de fraudes, la cybersécurité par la détection de comportements atypiques, la gestion de la relation clients et l'identification de clients mécontents ou risquant de se désabonner, l'étude d'implantation de commerces, la recommandation de produits et services, la segmentation de clients, l'analyse d'études cliniques ou de pandémies ou encore la génomique statistique.



Le data mining exploite des méthodes qui relèvent presque toutes du machine learning que nous avons déjà évoquées dans cet ebook : la classification pour labelliser des objets et situations, les régressions pour prédire des valeurs chiffrées, le clustering pour détecter des anomalies ou nouvelles classes d'objets et la découverte de corrélations entre données (qui relève de la PCA vue précédemment)[620]. Certaines de ces méthodes servent à analyser les données du passé et d'autres permettent de faire des prévisions[621].

Le concept de big data est alors associé à un grand nombre de technologies émergentes : les outils d'intégration de données, le format XML, les moteurs de recherche d'entreprise, les architectures distribuées, les gestionnaires de métadonnées, et les débuts du cloud. Bref, presque toute l'informatique d'entreprise et de l'Internet y passe !

Données

Les données gérées par le big data tournent essentiellement autour de :

- Bases de **clés-valeurs** (KVP =Key-value pairs) qui sont des tables simples à deux colonnes, une clé et une valeur contenant une donnée de n'importe quel type.
- Bases de données orientée **colonnes** qui servent surtout à gérer des données temporelles issues de capteurs divers. Exemples : Google BigTable, HBase, Cassandra et Hypertable.
- Des bases de données orientées **documents** dont le contenu est géré dans un format structuré comme JSON ou XML. Exemples : Apache CouchDB et MongoDB.

---

[620] Voir Council Post: How Is Big Data Analytics Using Machine Learning? par Chithrai Mani, octobre 2020.

[621] Source d'inspiration du schéma : Big Data Analytics Learning Lab 1 UN Data Innovation Lab 4, University of Nairobi, mars 2017 (84 slides).

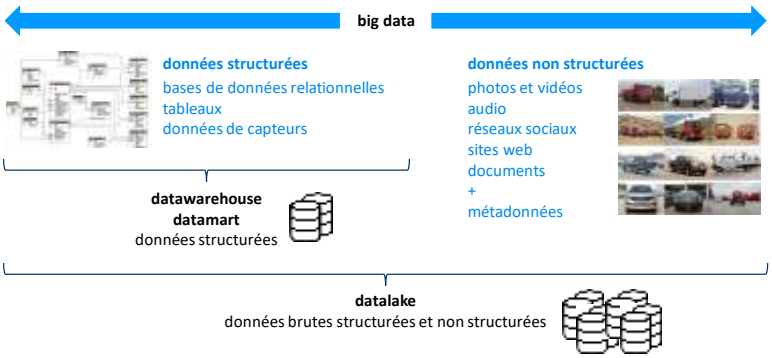
- Des bases de données orientées **graphes** qui gèrent des relations entre nœuds, liens et propriétés. Les bases telles que Cassandra qui savent gérer des jointures récursives peuvent faire l'affaire. Exemples : Infinite Graph et Neo4j.

Dans les architectures de big data, on rapproche les traitements des données de manière distribuée, à savoir que l'on distribue les traitements là où sont les données en évitant tant que possible de les déplacer. Dans le HPC, on fait plutôt l'inverse en chargeant un aussi gros volume que possible de données en mémoire vive pour y réaliser les calculs, souvent des calculs de simulation et scientifiques[622].

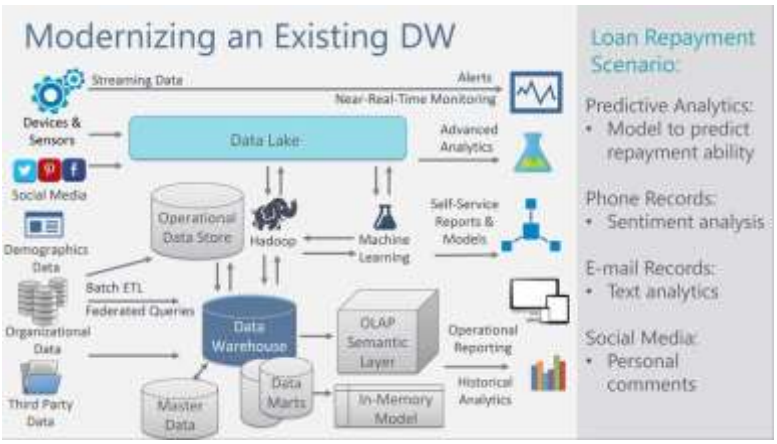
Data lakes

Dans le big data, les données sont généralement consolidées dans un **data lakes**[623], qui joue le rôle d'entrepot de données généraliste de l'entreprise.

Cela simplifie leur exploitation par les applications, notamment celles du machine learning et du deep learning qui peuvent exploiter plusieurs sources de données différentes et complémentaires. Il alimente aussi les datawarehouses (entrepôts de données) qui gèrent les données structurées intégrées dans des bases SQL.



Un data lake contient des données brutes tandis qu'un entrepôt de données classique contient des données déjà traitées et agrégées. Les sources des données peuvent être multiples. Les données sont structurées (bases relationnelles SQL), faiblement structurées (CSV, logs, données issues de capteurs et objets connectés) et non structurées (documents, mails, images). Les données peuvent être conservées pour être utilisées plus tard.



On ne connaît pas a priori leur utilisation lorsqu'on les accumule. Ce qui suppose tout de même que l'on en a tout de même une petite idée !

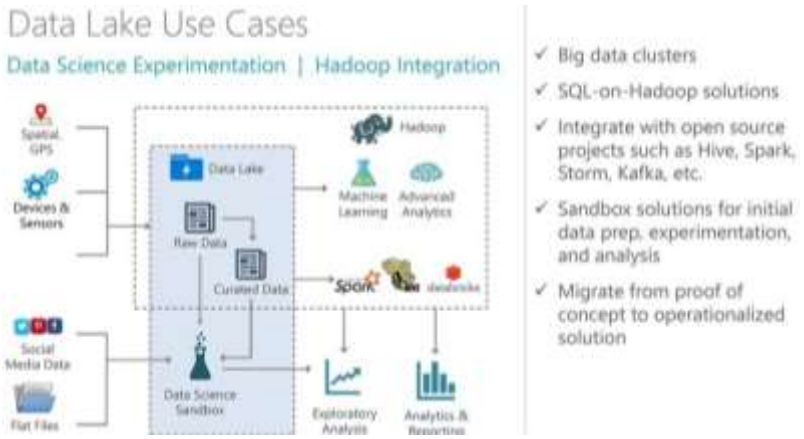
Elles sont physiquement stockées dans l'entreprise ou dans le cloud dans des bases de données distribuées (comme avec HDFS, Hadoop Distributed File Storage), dans des bases NoSQL ou dans des bases objets telles qu'Amazon S3 ou Azure Blog Storage.

---

[622] C'est bien expliqué dans Big Data : Informatique pour les données et calculs massifs - 5 – Technologies d'Hadoop par Stéphane Vialle, 2019 (44 slides). Voir l'ensemble des cours sur le big data de Stéphane Vialle de CentraleSupelec.

[623] Le concept de data lake a été évoqué pour la première fois par Dorian Pyle en 1999 dans Data preparation and data mining (466 pages) et promu ensuite en 2010 par James Dixon de la société américaine Pentaho. Voir Pentaho, Hadoop, and Data Lakes de James Dixon.

Un datalake est construit avec des outils d'ingestion de données, qui extraient données et métadonnées des bases qui l'alimentent, des outils de stockage souvent distribués et des moyens de recherche et d'extraction de données qui interrogent les bases avec des outils relationnels ou « no SQL ». Le tout n'est évidemment accessible que via une couche de sécurité et de droits d'accès.



Les data lakes sont utilisés pour l'analyse de données en mode lecture par opposition aux systèmes transactionnels qui fonctionnent en lecture/écriture et en temps réel[624]. Les données extraites du data lake doivent être préparées et netoyées avant d'alimenter l'entraînement de solutions de machine learning.

Un data lake « connaissance des clients » va ainsi conserver les données issus de systèmes disparates : CRM, SFA, centres d'appels, support technique et réseaux sociaux.

Elles peuvent être exploitées par **Spark** de la fondation Apache, qui fonctionne en mode distribué sur des données stockées en RAM dans les serveurs, ce qui permet d'entraîner des modèles de machine learning très rapidement via sa bibliothèque **SparkML**[625].

D'autres bibliothèques supportent le développement d'applications de machine learning distribué, comme **Mlbase**[626], **GraphLab** et **TensorFlow**.

Les infrastructures de big data profitent de divers progrès dans le stockage, le parallélisme et la communication entre serveurs. D'ailleurs, le marché du big data et de l'analytics a rapidement sauté dans le bain de l'IA, même si cela comporte probablement pas mal d'IA washing. L'offre est en tout cas plus qu'abondante.

Certains traitements restent allergiques au calcul distribué traditionnel en cluster. Par exemple dans le cas du deep learning, notamment appliqué à la reconnaissance d'images, il sera fortement recommandé de répartir les traitements localement via des pools de cartes graphiques embarquées sur les mêmes machines pour minimiser les échanges réseau. Ces derniers étant susceptibles de faire exploser les temps d'exécution.

Le champ de la gestion des données entre big data, data lake, cloud et machine learning est couvert par une offre des grands acteurs de l'IT (Oracle, Microsoft, IBM, …) ainsi que de nombreuses startups. En voici quelques-unes de notables :

- **Bigstream** (2015, USA, $5,5M) fournit une plateforme matérielle et logicielle pour gérer l'extraction de donnée, leur intégration et leur analyse. Leur Hyper-acceleration Layer permet de faire tourner les différentes briques logicielles, dont TensorFlow sur les meilleures plateformes matérielles (CPU, GPU, FPGA) pour optimiser leur performance. Ils ciblent les marchés de la publicité et de la finance.

[624] Informations et schéma issus de Designing a Modern Data Warehouse + Data Lake de Melissa Coates, 2017 (72 slides). Voir aussi ce cours de Omar Boussaid de l'Université de Lyon, Du Data Warehouse au… Data Lake, 2017 (54 slides).

[625] Voir Data Lakes A Solution or a new Challenge for Big Data Integration de Christoph Quix de l'Institut Fraunhofer, 2016 (28 slides) et Azure Data Lake What Why and How de Melissa Coates, septembre 2018 (81 slides).

[626] Voir MLbase: A Distributed Machine-learning System (7 pages).

- **Databricks** (2013, USA, $247M) propose sa plateforme Mlflow qui gère l'administration de ses processus de machine learning. C'est une sorte d'outil de gestion du workflow de ses processus de ML, avec l'ingestion de données, les transformations, le data cleaning et l'exploitation d'algorithmes de machine learning. L'offre de Databricks est intégrée dans Microsoft Azure depuis 2017.
- **Saagie** (2013, France, $35,1M) propose une plateforme open source de gestion de ses données dans le cloud utilisables notamment par vos applications de machine learning. C'est du backoffice qui gère les flux de données de l'entreprise, dont son data lake, gérant les accès utilisateurs, les processus d'ingestion de données, le respect du RGPD et la supervision. La startup vise en premier le marché de la banque et de l'assurance (50% des clients), puis l'industrie (25% comme chez Valourec) et le secteur public (25%). La startup emploie une centaine de personnes et a des bureaux de ventes au Royaume-Uni et aux USA.

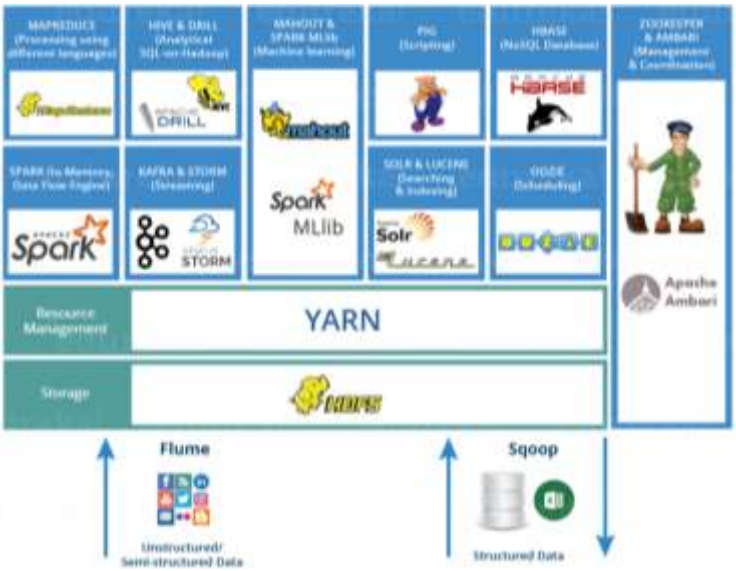
- **Sentient Technologies** (2007, USA, $174M) développe une solution d'IA massivement distribuable sur des millions de CPUs, visant les marchés de la santé, de la détection de fraudes et du du e-commerce. La société dit employer des méthodes d'IA avancées pour détecter des tendances dans les données. C'est de la "big data" revisitée. Le système imite les processus biologiques pour faire de l'auto-apprentissage. On trouve des morceaux de deep learning et des agents intelligents dedans. Ces agents sont évalués avec des jeux de tests et les meilleurs conservés tandis que les plus mauvais sont éliminés. Bref, c'est une sorte de Skynet. L'un des fondateurs de la société est français, Antoine Blondeau, et basé à Hong Kong.
- **Algorithmia** (2013, USA, $12,9M) propose son outil logiciel AI Layer pour déployer dans le cloud et de manière répartie des modèles entraînés de machine learning. Le tout est associé à une bibliothèque de plus de 4000 micro-services intégrables dans les applications. Il en résulte un modèle déployé dans le cloud exploitable sous forme de service. L'outil fait gagner du temps aux data scientists et automatise la partie « dev ops » (les opérations de déploiement des applications à la croisée des chemins entre les développeurs et les responsables de l'infrastructure IT) du déploiement du projet.
- **Tamr** (2014, USA, $59,2M) est une autre startup qui propose des outils d'ingestion de données disparates pour les applications de big data et de machine learning qui exploite en amont de l'expertise humaine, donc pas scalable, pour le nettoyage des données. La startup exploite des recherches du MIT Computer Science and Information Lab (CSAIL).
- **Skymind** (2014, USA, $6,3M) est une autre startup qui facilite le déploiement de solutions de machine learning dans le cloud avec sa Skymind Intelligence Layer (SKIL). Elle est notamment adaptée à l'entraînement incrémental de modèles de machine learning. L'outil est surtout destiné aux équipes informatiques qui déploient les applications. La startup est aussi à l'origine de Deeplearning4j.org, un framework open source de deep learning pour Java, distribué sur Apache Spark et Hadoop, écrit en C et C++ et CUDA pour l'accès aux GPU Nvidia.

- **Ayasdi** (2008, USA, $106M) interprète aussi de gros volumes de données pour y identifier des signaux faibles pertinents. Le projet a démarré à Stanford et avec des financements de la DARPA et de la NSF, l'équivalent américain de l'Agence Nationale de la Recherche française.
- **Versive** (2012, $57M), anciennement Context Relevant, propose des outils d'analyse prédictive applicables à différents marchés. Le glissement sémantique semble généralisé : au lieu de parler de big data, ce qui est trop vague, les startups parlent plutôt d'analyse prédictive qui exploite de gros volumes de données.

Hadoop

Parmi les outils logiciels du big data, le plus emblématique est probablement **Hadoop** qui sert à répartir des traitements de données peu ou pas structurées[627] sur des clusters de serveurs dits « de commodité » et en mode batch. C'est la base des architectures distribuées du big data. Son histoire est riche. Ce projet open source est aujourd'hui sous l'ombrelle de l'Apache Foundation. Sa genèse remonte à la création du Google File System en 2003 ainsi que de MapReduce en 2004.

Le projet Hadoop a pris forme en 2006. L'appellation Hadoop est celle d'un jouet utilisé par l'enfant de Doug Cutting, l'un de ses créateurs qui travaillait alors chez Yahoo! Hadoop est issu du logiciel de crawling du web **Apache Nutch** lancé en 2002, lui-même faisant partie du projet de moteur de recherche **Lucene**. Yahoo! en a été l'un des premiers utilisateurs en 2008 pour générer l'index de son moteur de recherche. En 2009, il était alors capable de trier un To de données en une minute, sur un cluster de 910 nœuds.



À ce jour, Hadoop en est à sa version 2.10. Il comprend les briques suivantes :

- **HDFS** (Hadoop Distributed File System), un système de gestion distribué de gros fichiers. Les clusters de serveurs peuvent comprendre plusieurs milliers de machines. Certaines de ces machines gèrent l'accès au stockage et d'autres, les traitements. Les nœuds de HDFS gèrent des sortes des tables d'allocation de fichiers qui savent où les blocs qui les composent sont stockés physiquement et de manière redondante dans au moins trois endroits différents[628]. Les tables d'allocation sont dans un serveur 'NameNode' qui gère les métadonnées décrivant les fichiers et les données dans des 'DataNodes'. Le tout constitue un cluster Hadoop avec en général un serveur pour le NameNode et plusieurs serveurs, jusqu'à des milliers, pour les DataNodes qui lui sont associés. Les données du NameNode sont elles-mêmes répliquées, depuis la version 2.0 de HDFS[629]. Hadoop utilise aussi des JournalNodes qui conservent les logs des modifications de NameNode pour faire de la restauration de système en cas de panne.

---

[627] Par opposition aux données structurées des bases de données relationnelles. Mais on peut exploiter les données de bases de données relationnelles SQL distribuées en s'appuyant sur les briques logicielles de Hadoop.

[628] Il est possible de se passer de cette redondance coûteuse en espace disque et de la remplacer par du stockage sur disques RAID où l'overhead en stockage de données est plus faible, grâce à l'emploi de bits de parité, les données étant réparties sur plusieurs disques, au moins trois pour le RAID 5.

[629] Dans les offres en cloud, HDFS peut être remplacé par des architectures de stockage propriétaires comme Amazon S3, Google Cloud Storage et Microsoft Azure Storage.

- **Hadoop Common,** commun à tous les modules, qui comprend les bibliothèques et utilitaires communs aux différents modules d'Hadoop.
- **YARN** (Yet Another Resource Negotiator), un framework de gestion des tâches et des ressources des clusters de serveurs. Il s'intercale entre HDFS et MapReduce.
- **MapReduce**, un framework qui sert à paralléliser les traitements de données sur plusieurs serveurs dans des clusters. Il est basé sur YARN qui parallélise les traitements de gros volumes de données. Pourquoi MapReduce ? Map pour la distribution des traitements et Reduce pour leur consolidation sous forme de résultat exploitable.
- **Ozone** : un gestionnaire de stockage distribué d'objets qui s'appuie sur YARN et HDFS. Il présente la particularité de gérer le RGPD et en particulier le droit à l'oubli par la destruction automatique des donnés liées à des métadonnées que l'on supprime.

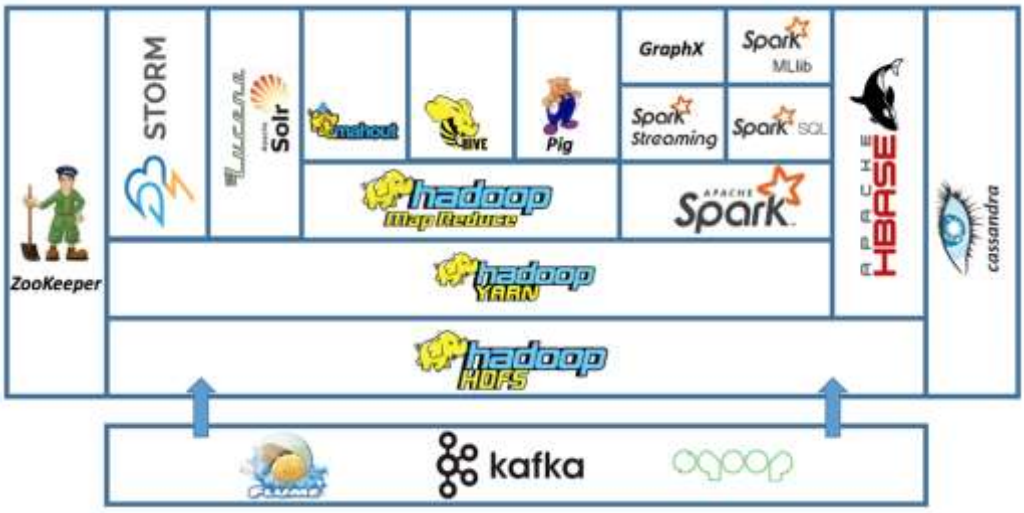
Hadoop est aussi la base d'un grand nombre de projets open source de la fondation Apache avec en particulier[630] :

- **Pig**, un langage de scripting pour manipuler les données distribuées.
- Des outils de gestion de données avec deux bases NoSQL distribuées : **Cassandra** et **HBase** (pour les très grandes tables), **Hive**, un gestionnaire de requêtage en langage SQL (HiveQL) pour des batchs longs de data mining, **Drill**, un outil de requêtage de données structurées et non structurées pour l'exploration de données et les business analytics et **Giraph** pour le requêtage de graphes (calcul de distances, etc) comme ceux de la base de graphes **TinkerPop**.
- Des briques logicielles pour le machine learning : **Mahout** (bibliothèque de machine learning et de data mining avec fonctions de clustering, de classification, de recommandation, d'analyse de textes), **Spark** (data analytics exploité par le moteur de machine learning **MLlib**, et qui accède directement à HDFS sans passer par MapReduce) et **Submarine** (distribue des traitements de machine learning et de deep learning dans des clusters distribués, en particulier à base de GPU de type Nvidia). Ces briques associent les algorithmes classiques de machine learning à l'exploitation de Hadoop pour en répartir les traitements dans des architectures distribuées à même de traiter de gros volumes de données d'entraînement. Le data mining du big data est en fait quasiment synonyme de machine learning si l'on s'en tient au contenu de la littérature technique sur le sujet ! Il s'applique aussi bien aux modèles d'entraînement supervisés que non supervisés. Il consomme des données faiblement structurées de type clé-valeur ou colonnes issues souvent de bases NoSQL. Comme nous l'avons vu dans les parties précédentes, l'entraînement des modèles de machine learning est celui qui consomme le plus de ressources de calcul et d'accès aux données. C'est aussi le cas lorsque celui-ci s'appuie sur de gros volumes de données. Les modèles associant machine learning et big data doivent de plus intégrer le fait que les données d'entraînement sont très dynamiques. Cela pousse à adopter des solutions d'entraînement « incrémentales » des modèles de machine learning et de deep learning, comme l'apprentissage par transfert[631].
- **Storm**, qui s'appuie sur YARN et fournit un système de traitement de données de flux en temps-réel. Cela complète bien les autres briques de Hadoop qui sont plutôt adaptées pour des traitements en batch.
- **Oozie**, un outil de planification de tâches à base de workflows (scheduling).
- **Solr**, un moteur de recherche d'entreprise s'appuyant sur le projet Lucene.

---

[630] Source du schéma : https://www.edureka.co/blog/hadoop-ecosystem.

[631] Source du schéma suivant : 21 Must-Know Open Source Tools for Machine Learning you Probably Aren't Using (but should!) par Mohd Sanad Zako Rizvi, juillet 2019.



- **Ambari** : un outil web de gestion et de monitoring de clusters Hadoop.

Une partie de ces outils est packagée dans **Apache BigTop**, qui est un peu l'équivalent de XAMPP (briques Apache/PHP/MySQL pour se créer un serveur LAMP) pour la création d'un serveur de big data de test.

Autres outils

Si la fondation Apache joue un rôle central dans le développement des outils du big data, elle n'est pas seule. À commencer par les sociétés privées qui repackagent et complètent Hadoop avec des briques logicielles plus ou moins propriétaires.

Elles sont à Hadoop ce que Red Hat est à Linux. Il faut notamment compter avec **Cloudera** (2008, USA, $1B) où Doug Cutting officie comme architecte en chef et qui a fait l'acquisition de **Hortonworks** (2011, USA, $248M) en 2018, **HPE** (notamment via son acquisition de **MapR Technologies** en 2019[632]), **IBM** et **EMC**.

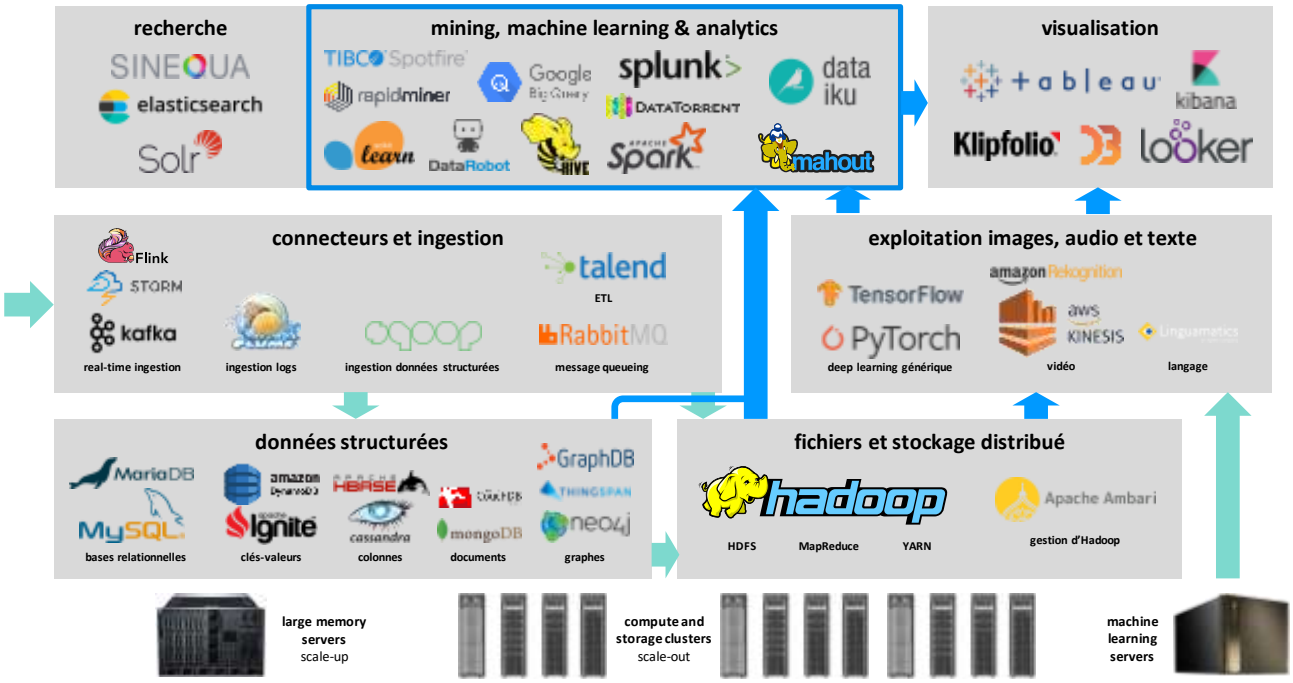
Il faut bien entendu y ajouter **Amazon AWS** (avec Elastic MapReduce (EMR), Elastic Compute Cloud (EC2) et Simple Storage Service (S3)) et **Microsoft Azure** qui repackagent et complètent aussi les briques d'Hadoop dans leurs offres de cloud.

La panoplie des outils du big data est aussi complétée par :

- Des outils d'**ETL** (Extract Transform & Load) pour convertir et formater les données à charger dans Hadoop. Ceux du français **Talend** en font partie. La fondation Apache gère de son côté le projet **NiFi** qui provient de la NSA.
- Des bases de données pour les « operational analytics » temps-réel comme **Rockset** (2016, USA, $61,5M) ou **Scuba** (2013) qui est utilisé en interne chez Facebook.
- Des outils divers dédiés au traitement du **langage** et des **images**.

[632] MapR couvre à la fois les besoins en gestion de datalakes et ceux du stockage.

- Des outils de **traitement de graphes** comme Pregel, Ligra, GraphChi, Xstream et GridGraph. Ils sont particulièrement utilisés dans les réseaux sociaux.
- Divers **outils de visualisation** comme ceux d'IBM, SaaS, de Tableau (acquis par SalesForceen 2019) et de nombreuses startups.



Il faut noter que les grandes bases de données relationnelles utilisées dans les applications transactionnelles sont généralement situées hors du champ du big data. Elles sont gérées à part, éventuellement associées à des data warehouses (entrepôts de données) et des datamarts (magasins de données, sous-ensembles des entrepôts de données) pour leur composante décisionnelle et analytique. Voici pour résumer tout cela un schéma qui positionne et structure assez bien le jargon du big data[633].

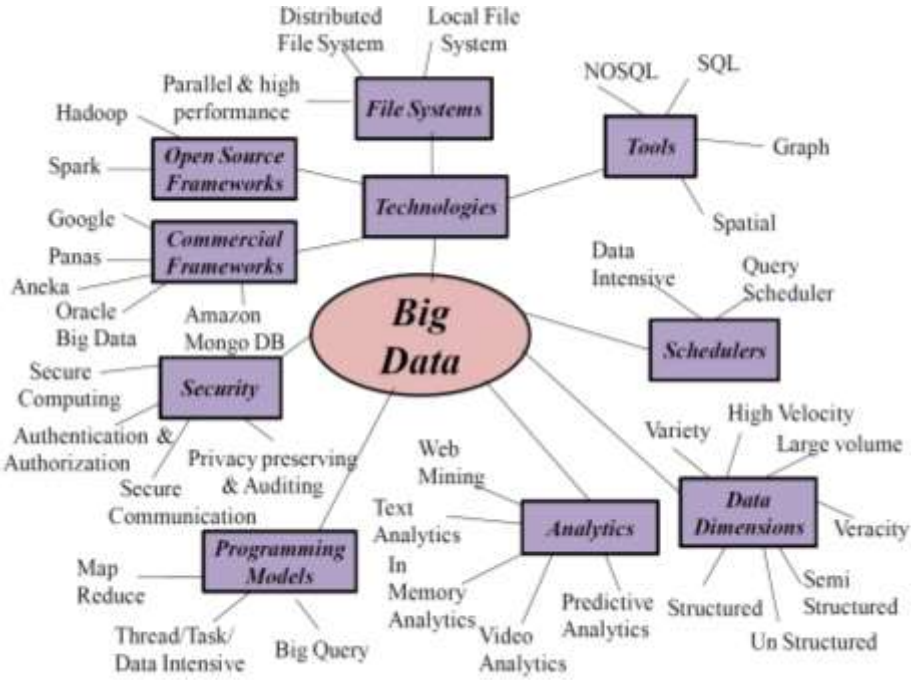


Figure 4. Big Data Taxonomy

[633] Voir The Anatomy of Big Data Computing par Raghavendra Kune et al, 2015 (33 pages).

D'autres sources d'information permettent de se faire une idée de la manière dont le big data est déployé dans les grands acteurs, particulièrement de l'Internet[634].

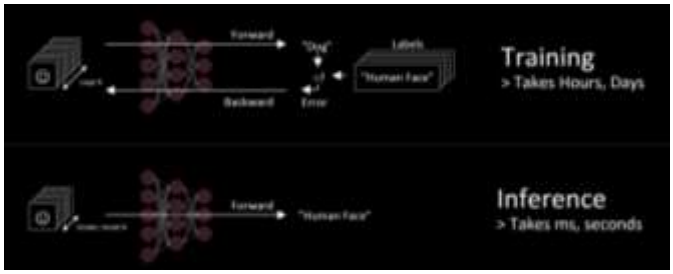
Il n'existe pas d'organisation unique type. Chaque entreprise ou acteur de l'Internet fait ses choix d'urbanisation de système d'information en fonction de critères très divers comme les données exploitées, les exigences de temps de réponse des applications, les usages du machine learning, les contraintes de l'existant (le « legacy ») ainsi que les choix d'opérateurs de cloud ou d'implantation de serveurs « on-premise » (dans les salles blanches appartenant aux entreprises et pas à des opérateurs de cloud).

Cloud

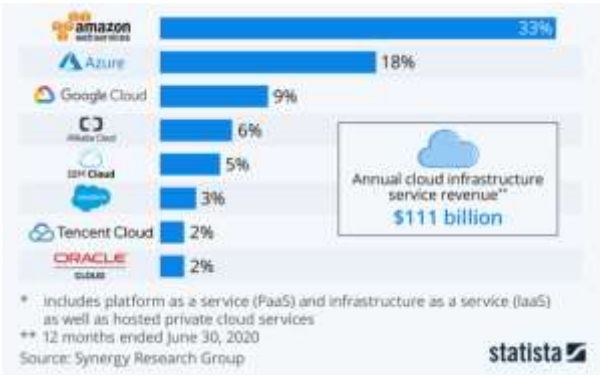
Les applications de l'IA font aussi appel aux ressources du cloud, en particulier dans les phases d'entraînement et surtout pour les startups qui ne peuvent pas disposer de leur propre data center.

Les grandes entreprises auront à gérer un équilibre entre leurs data-center « on premise » (chez elles) et dans des clouds privés et publics. La rapidité d'évolution des technologies de processeurs neuromorphiques et GPU que nous avons vues plus haut justifie le choix du cloud pour éviter l'obsolescence rapide de ses propres infrastructures.

Les infrastructures en cloud doivent pouvoir « scaler » pour s'adapter à l'entraînement de modèles de machine learning et deep learning nécessitant d'aligner parfois des milliers de serveurs. Une fois les modèles entraînés, leurs besoins en ressource machine sont plus faibles, surtout pour les solutions de deep learning[635].



Ce n'est pas pour rien, par exemple, qu'un GPU Nvidia ou un Google TPU offre une puissance de calcul située aux alentours de 100 TFLOPS (opérations en nombres flottants) tandis que les unités de traitement neuronales embarquées dans les smartphones les plus récents comme le Huawei Pmate 40 et l'iPhone 12 ont une puissance de calcul située entre 10 et 15 TOPS (opérations en nombres entiers) ! L'exécution d'un réseau de neurones est bien plus rapide que son entraînement ! La plupart des offres de cloud intègrent maintenant la capacité à entraîner et exécuter des modèles de machine learning et de deep learning.



Google	**Google Cloud Machine Learning**
amazon	**Amazon Artificial Intelligence & Alexa**
Microsoft	**Microsoft Cognitive Services + Azure**
IBM	**IBM Watson Cloud Servers**
OVH.com	**hébergement de serveurs Nvidia DGX**
clever cloud	**Clever Grid et cartes Nvidia GT1070**
Scaleway	**hébergement de serveurs Nvidia DGX**

[634] Voir Big data application architecture Q&A par Nitin Sawant et Himanshu Shah, 2015 (157 pages), Exploring the boundaries of big data par Bart van der Sloot et al, 2020 (288 pages) et How to build and run a big data platform in the 21st century par Ali Dasdan, 2019 (133 slides). Enfin, Big data in practice - how successfull companies use big data analytics to deliver extraordinary results par Bernard Marr, 2016 (323 pages) est un compendium d'étude de cas de secteurs divers (Wallmart, CERN, Netflix, Shell, Facebook, John Deere, LinkedIn, BBC, Palantir, Airbnb, Ralph Lauren, etc) mais sans qu'elles soient très bien documentées, notamment sur les aspects techniques et les architectures utilisées.

[635] Source du schéma : Edge Intelligence : Convergence of edge compute and machine learning, AMD, Allen Rush, 2018 (30 slides).

C'est le cas chez **Google**, **IBM**, **Microsoft**, **Amazon**[636], **Paperspace**[637] et même **OVH** qui dispose notamment de serveurs DGX de Nvidia et propose une offre OVHCLoud AI comprenant une place de marché de modèles de machine learning prêts à l'emploi, issus d'une trentaine de partenaires[638].

On compte aussi des acteurs locaux tels que **Scaleway** (filiale de free, anciennement Online.net) et **CleverCloud**. Le premier lançait en mars 2019 une offre d'instances cloud de GPU Nvidia V100 couplées à des Intel Xeon Phi à des tarifs compétitifs de 1€/heure ou 500€/mois. Le second, avec des Nvidia GTX 1070[639].

Sachant qu'il est bon de vérifier pour les applications les plus critiques que les données restent en France ou en Europe. Il existe des ressources en cloud originales comme celles d'**iExec** (2016, France, $12M) une startup issue de travaux de recherche du CNRS et d'Inria qui propose une architecture répartie de traitements décentralisée via une blockchain Ethereum.

Les utilisateurs de serveurs et de PC peuvent ainsi louer leur puissance machine. Cela suppose que les entraînements et inférences ont été préparés pour pouvoir être distribués sur des machines disparates[640]. L'IA peut aussi servir à gérer des infrastructures en cloud comme le fait maintenant Google pour réduire sa consommation d'énergie[641].

Pour réduire les transferts de données entre les serveurs locaux et les serveurs nécessaires à l'apprentissage, l'approche du **Federated Learning** consiste à entraîner des réseaux neuronaux localement et à ne transmettre que les propriétés et paramètres des réseaux neuronaux locaux à un réseau neuronal d'un niveau supérieur. Cela peut servir à éviter d'exposer ses données à l'extérieur, en lieu et place ou en complément du chiffrement homomorphe.

Cependant, pour ce qui est des GAFAMI, l'usage de leurs ressources en cloud pour héberger des applications métiers et leurs données ne signifie pas que celles-ci serviront à améliorer les frameworks de ces géants ni à entraîner leurs propres IA, sauf à violer les termes des contrats clients/fournisseurs associés à ces services d'hébergement où à être victimes de failles de sécurité involontaires chez ces opérateurs[642].

L'exploitation des ressources du cloud pour l'IA est une discipline en soi avec ses nombreuses composantes technologiques que nous n'étudieront pas en détail : les kubernettes qui facilitent les déploiements, les micro-services qui modularisent les applications, l'exploitation de serveurs adaptés aux applications associant CPU, GPU et TPU, etc.

---

[636] Voir Comparing Machine Learning as a Service: Amazon, Microsoft Azure, Google Cloud AI, IBM Watson par Olexander Kolisnykov, août 2019. Voir aussi la soluton AI-Compare de la startup française DataGenius. Voir AWS to offer NVIDIA A100 Tensor Core GPU-based Amazon EC2 instances | Amazon Web Services par Geof Murase, mai 2020.

[637] Paperspace (2014, USA, $23M) est un challenger américain des leaders du cloud, issu du cloud gaming et du « desktop as a service ». Dans leur offre Gradient, ils proposent des batteries de GPU en cloud pour les besoins d'applications de machine et deep learning. L'offre s'adresse aux devops et aux data scientists. Ils s'appuient sur des outils open source tels que Fast.AI, H2O Driverless AI et TensorFlow de Google. Le tout est géré à partir de l'interface web de programmation Jupyter Notebook. Il permet des déploiements sur AWS ou le cloude de Google, ou dans des containers Docker sur les architectures Kubernetes. Voir Qui est Paperspace, ce challenger de l'IA en mode cloud ? par Alain Clapaud, janvier 2019.

[638] Voir https://www.ovhcloud.com/en-ie/public-cloud/ai-training.

[639] Voir Clever Cloud launches GPU-based instances par Romain Dillet, juillet 2019.

[640] Voir How Blockchain And AI Can Help Master Data Management par Naveen Joshi, 2019.

[641] Voir Artificial intelligence is now directly controlling cooling at Google data centers par Abner Li, août 2018.

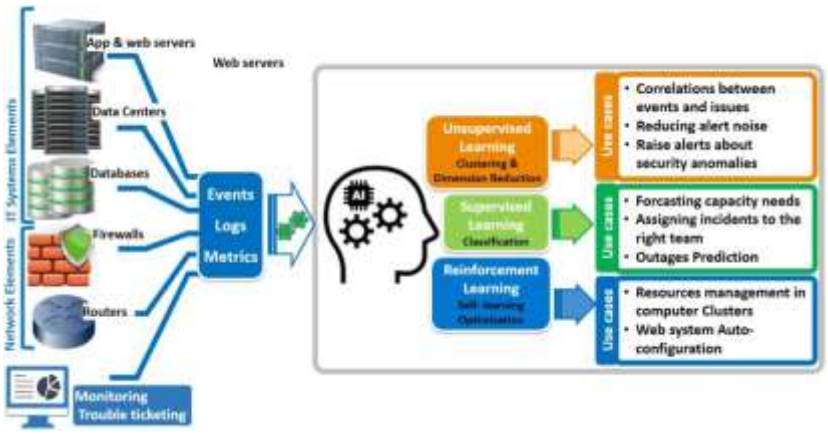
[642] Voir par exemple L'accord controversé de Google avec plus de cent cinquante hôpitaux aux Etats-Unis par Alexandre Piquart, novembre 2019. Ce partenariat avec Ascension lui donnerait accès aux données médicales de millions de patients sans leur consentement. En apparence, c'est une vente de services de cloud pour cette chaîne d'hôpitaux. Ascension comprend 2600 sites de soins et 150 hôpitaux et cinquante équivalents d'EPAHD. Les applications de ce client hébergées chez Google vont gérer les dossiers médicaux de ses patients. Il fera appel à des briques d'IA de Google. Mais ces données devraient être gérées par Ascension même si elles sont hébergées chez Google. Elles pourraient cependant alimenter les services de ce dernier ou des filiales santé d'Alphabet, notamment dans le cadre du projet Project Nightingale. La communication de Google a été très ambigue sur ce point. Et une enquête fédérale a été lancée aux USA. Voir Google, Ascension data partnership sparks federal probe par Jessica Kim Cohen, novembre 2019.

AIOps

Les outils de gestion des opérations de l'infrastructure IT ont aussi adopté l'IA et en particulier le machine learning. Et pour cause puisqu'ils manipulent de gros volumes de données qui peuvent être exploités pour détecter des anomalies et faire des prévisions.

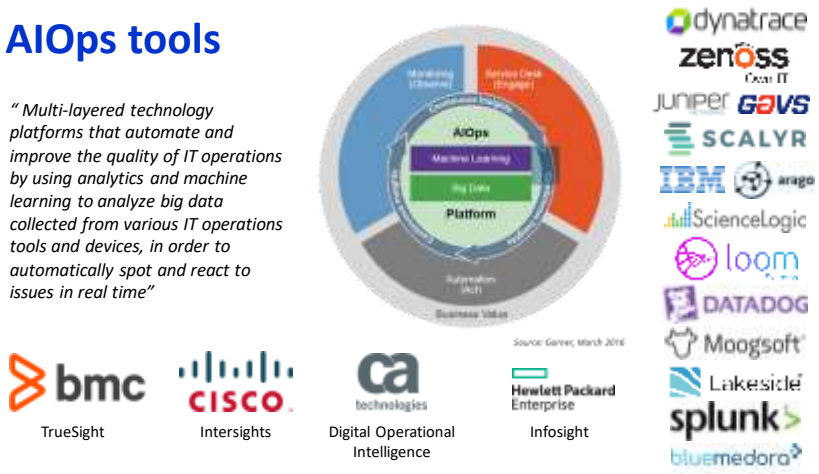
Ces outils exploitent les trois formes de machine learning : supervisé (avec des données labellisées), non supervisé (avec des données non labellisées, pour détecter des anomalies) et par renforcement (pour la gestion assistée de configurations)[643]. L'accent est de plus en plus mis sur la maintenant prédictive des infrastructures pour détecter les défaillances le plus en mont de leur émergence[644].

Ces outils d'AIOps s'alimentent de données variées issues des serveurs d'applications, serveurs de données, bases de données, pare-feux, routeurs réseaux et bases d'incidents[645].



Les acteurs de ce marché sont les fournisseurs habituels d'outils de supervision d'infrastructure (**Bmc**, **Cisco**, **CA**, **HPE** [646]) et un grand nombre de startups émergentes.

On compte notamment **Zenoss** (2005, USA, $45M) qui prédit les ruptures de services, **Gavs-tech** (1998, Inde), **Scalyr** (2011, USA, $27,6M) qui gère les logs et les alertes, **ScienceLogic** (2003, USA, $109M) qui monitore les ressources hybrids de l'IT (on premise & dans le cloud), **Loom Systems** (2015, USA, $6M), **Moogsoft** (2011, USA/UK, $90M), **Datadog** (2010, USA, $147M).



Et puis aussi **Lakeside Software** (1997, USA), **Splunk** (2004, USA), **Arago** (1995, Allemagne, $55M) et **Packet AI** (2018, France) qui détecte et résout les incidents dans les infrastructures. IT. **Lenovo** propose aussi LiCO, un outil d'administration des ressources de l'IA dans le cloud[647].

Dynatrace (2005, Autriche) est un acteur des AIops qui fournit une solution tout en un de monitoring de ses ressources et applications : suivi de performance d'applications déployées en cloud hybride, suivi d'infrastructure et business analytics, notamment via Davis, leur moteur de machine

---

[643] Voir Machine learning and IT infrastructure management automation (AIOps) par Youssef Fenjiro, 2018 et aussi Glossaire IT #3 Intelligence artificielle, 2019 (51 pages).

[644] Voir How To Rise Above The ITOps Chaos Using AI par Andy Thurai, février 2020.

[645] Source du schéma : TrueSight Platform for AI Ops Elevate IT Ops with Artificial Intelligence, 2018 (10 slides).

[646] L'offre d'AIops d'HPE,InfoSight, est issue de l'acquisition de Nimble Storage (2008, USA, $98,8M) en 2017, qui était spécialisé-dans les systèmes de stockage à base de SSD. Cela explique que InfoSight soit une brique d'AIops focalisée sur la surveillance des systèmes de stockage.

[647] Voir LiCO: Simplifying AI Development, 2018 (11 pages).

learning. En septembre 2020, ils annonçaient intégrer leur plateforme de monitoring avec les services de télémétrie Microsoft Azure Monitor couvrant 80 services du cloud Azure.

On peut aussi citer **Juniper** qui, avec l'acquisition de **Mist** (2014, USA, $104M) en 2019, intègre maintenant des fonctions d'AIops dans ses outils de monitoring de réseaux comme son Marvis Virtual Network Assistant qui est exploitable via des requêtes exprimées en langage naturel. Et puis **IBM** Watson AIops lancé en mai 2020 et qui s'appuie sur les capacités de machine learning de Netcool Operations Insight et sur la plateforme de conteneurs Red Hat OpenShift.

Enfin, citons ce projet de gestionnaire de batchs de HPC entraîné par appentissage par renforcement qui sert à optimiser l'usage des ressources et le temps de réponse du supercalculateur. C'est un travail de recherche mené pour les centres de calcul du Département de l'Energie US[648].

Robotic Process Automation

Depuis 2015, la **Robotic Process Automation** décrit les outils d'automatisation des processus internes des entreprises couvrant la finance et le marketing.

Elle consiste à permettre à des agents à base d'AI de naviguer par eux-mêmes dans les différentes applications de l'entreprise afin de mener des tâches prédéfinies comme la collecte de documents[649].

L'IA permet en théorie à ces agents d'évoluer par eux-mêmes pour ingérer de nouvelles règles. Les études de cas se font jour depuis 2015[650]. On peut distinguer trois niveaux de RPA selon le rôle de l'IA dans l'automatisation. Le dernier niveau est aussi dénommé IPA, pour Intelligent Process Automation. Il se produit lorsque des briques d'IA sont utilisées comme la reconnaître d'images et

---

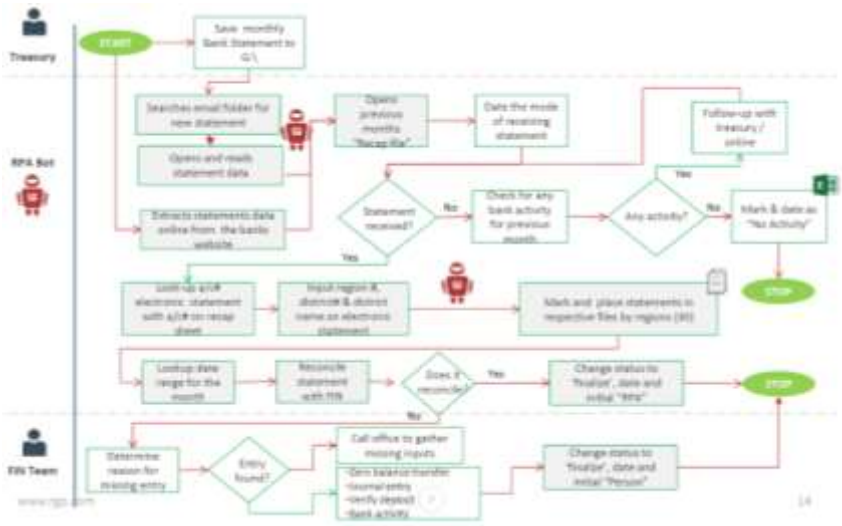
[648] Voir RLScheduler: An Automated HPC Batch Job Scheduler Using Reinforcement Learning par Di Zhang et al, octobre 2019 (14 pages).

[649] Voir How to choose between robotic process automation and back-end system automation 2018 (15 pages), Setting up a robotic process automation center of excellence 2013 (14 pages), Robotic Process Automation (RPA) Tutorial: What is, Tools & Example, un bon tutorial sur la RPA.

[650] Voir Robotics Process Automation 6 questions to master it, janvier 2017 de la société de conseil parisienne Ailancy.

de textes dans les documents (OCR), la reconnaissance vocale, les analyses sémantiques (NLP), celles des comportements et des sentiments, la détection de patterns avec le Machine Learning et l'interaction avec les utilisateurs via des agents conversationnels.

L'un des impacts de cette robotisation des processus sera de réduire l'emploi dans les entreprises concernées et surtout chez leurs sous-traitants, et notamment en Inde pour les entreprises anglo-saxonnes[651].



source : ROA The Impact on Internal Audit, RGP 2017

UiPath (2005, USA/Roumanie, $1,2B) est une licorne positionnée exactement sur ce créneau-là[652], fondée par un ancien développeur de Microsoft d'origine roumaine. La startup est actuellement le leader mondial de la RPA avec un CA 2019 de $360M. La filiale française est installée depuis 2017 après plus d'une cinquantaine de personnes.

La solution se découpe en trois parties : UiPath Studio qui permet de décrire les processus business de l'entreprise, et aussi d'enregistrer des sessions d'accès à des applications, UiPath Robot qui gère l'automatisation des processus et UiPath Orchestration qui permet de gérer le robot et l'orchestration (vidéo). Il va sans dire que cette automatisation des processus n'a rien de magique et qu'elle requiert beaucoup de paramétrage manuel. Cette forme de RPA reprend les anciens concepts des solutions de gestion de workflow. La solution de RPA s'intègre notamment avec les Amazon Web Services, Microsoft Azure, l'infrastructure de cloud d'Oracle et les applications SAP.

Derrière UIPath se battent en duel un bon nombre de sociétés difficiles à départager telles que **Automation Anywhere** (2003, USA, $840M)[653], **Kofax** (1991, USA, qui fait partie de Lexmark depuis 2015), **Blue Prism** (2001, UK, $59M), **Aisera** (2017, USA, $34,5M), **Catalytic** (2014, USA, $44M, qui automatise le travail de knowledge workers) et **Contextor** (2000, France, 600K€, acquis par SAP en 2018).

Les fournisseurs de solutions de RPA proposent maintenant des briques de découverte de processus existants (« process discovery capabilities » ou « advanced process mining »), notamment dans les

[651] Voir Introduction to Robotic Process Automation, a primer, de l'IRPA, Institute for Robotic Process Automation, 2015. Source du schéma : RPA and Beyond de TATA, juin 2017. Voir aussi The robots are ready. Are you? (28 pages), une étude de Deloitte sur l'impact de la RPA dans les entreprises.

[652] Voir Les "robots logiciels" de cette startup roumaine prennent le travail peu qualifié d'employés de bureau, août 2017. Certains considèrent que l'on est en pleine bulle de la RPA avec les surpromesses associées. Voir The Big RPA Bubble par Alan Trefler, décembre 2018.

[653] En août 2019, Automation Anywhere faisait l'acquisition de la startup parisienne Klevops, spécialisée dans les chatbots de relation client.

ERP maison[654]. On en trouve notamment chez **Kryon Systems** (2008, USA, $13M), **myInvenio** (2013, Italie) et **Abbyy** (1989, USA, $6M).

La RPA sophistiquée doit s'appuyer sur l'exploitation de données non structurées et notamment le langage. D'où l'intérêt de **Recital** (2017, France) qui propose une solution de recherche de texte qui cible la finance et l'assurance. Elle automatise le traitement des mails, extrait des données de documents textuels. Elle a Natixis comme premier client.

Parmi les grands acteurs, il faut compter avec **Microsoft** qui propose sa solution de RPA Power Automate depuis 2020, comprenant l'outil de création de processus UI Flows. **IBM** s'est aussi lancé dans la RPA en 2020, via l'acquisition de la startup **WDG Automation** (2014, Brésil) qui est spécialisée dans l'automatisation de processus clients via le traitement du langage.

Energie

Les solutions à base d'IA et notamment celles du deep learning sont nous l'avons vu de grandes consommatrices d'énergie, aussi bien pour l'entraînement des modèles que pour les inférences, surtout lorsque les utilisateurs sont nombreux comme avec les applications grand public, surtout alors que les modèles deviennent de plus en plus complexes[655]. Ainsi, l'entraînement du chatbot Meena de Google avec une base d'entraînement comprenant 40 milliards de mots utilise-t-il un système à base de 2048 processeurs TPU pendant 30 jours, consommant environ 300 MWh[656].

Cela alerte depuis quelques années, ce d'autant plus que le poids du numérique dans la consommation d'électricité augmente de manière galopante[657]. Dans le mix énergétique, les data centers sont avec les réseaux et télécommunications les premières sources de dépense énergétique[658].

Il existe plusieurs stratégies pour réduire l'empreinte énergétique de l'IA. La première que nous avons déjà évoquée dans cette partie sur les composants revient à créer des composants aussi optimisés que possible pour les traitements du machine learning.

La seconde consiste à déporter les traitements dans les objets connectés. Cela présente plusieurs avantages : les composants de l'embarqué ont en général une bien meilleure efficacité énergétique, celle-ci est amplifiée par la réduction des coûts télécoms et qui plus est, cela permet de mieux protéger la vie privée.

---

[654] Voir Driving Successful RPA Implementation with Automated Process Discovery par Ryan M. Raiker, 2019.

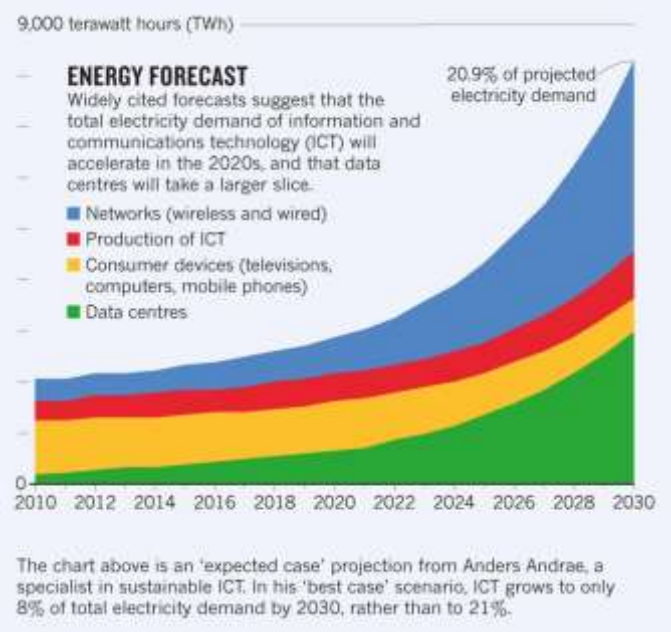
[655] Voir AI and Compute par OpenAI, mai 2018, qui évalue l'évolution galopante de la complexité des modèles de deep learning.

[656] Voir Towards a Human-like Open-Domain Chatbot par Daniel Adiwardana et al, février 2020 (38 pages).

[657] Voir AI Can Do Great Things—if It Doesn't Burn the Planet par Will Knight, janvier 2020 et AI Is an Energy-Guzzler. We Need to Re-Think Its Design, and Soon par Peter Rejcek, février 2020 et Deep Learning's Carbon Emissions Problem par Rob Toews, juin 2020.

[658] Voir How to stop data centres from gobbling up the world's electricity par Nicola Jones, Nature, septembre 2018. Qui fait référence à New perspectives on internet electricity use in 2030 et Supplementary Information for New perspectives on internet electricity use in 2030 par Anders Andrae, juillet 2020.

La troisième consiste à simplifier les calculs avec la quantisation qui revient à remplacer le calcul en nombres flottants par du calcul en nombres entiers. Enfin, il existe aussi des méthodes permettant de réduire la taille des jeux de données d'entraînement des modèles[659]. Certaines de ces méthodes sont loin d'être universelles et sont créées de manière ad-hoc cas par cas[660]. Ces méthodes sont évoquées dans l'initiative **Green.AI** qui vise à déployer de l'apprentissage d'IA faiblement énergétique[661]. Elle explique le lancement d'un outil tel que **CodeCarbon**, un logiciel open source qui sert à évaluer le bilan carbone des algorithmes d'intelligence artificielle, développé par le MILA conjointement avec le BCG[662].



Performance énergétique de l'IA

L'homme ne consomme en moyenne que 100 Watts au repos, dont 20 Watts pour le cerveau. C'est un excellent rendement. Tout du moins, pour ceux qui font travailler leur cerveau. Ce n'est pas facile à égaler avec une machine et pour réaliser les tâches de base que réalise un humain. Les supercalculateurs consomment au mieux quelques KW et certains dépassent les 10 MW.

Des progrès sont cependant notables dans les processeurs mobiles. Consommant moins de 5 W, ils agrègent une puissance de calcul de plus en plus impressionnante grâce à des architectures multi-cœurs, à un fonctionnement en basse tension, aux technologies CMOS les plus récentes comme le FinFET (transistors verticaux) et/ou FD-SOI (couche d'isolant en dioxyde de silicium réduisant les fuites de courant dans les transistors et améliorant leur rendement énergétique) et à une fréquence d'horloge raisonnable (entre 1 GHz et 2,5 GHz).

La technologie FD-SOI issue de STMicroelectronics et Soitec gagne petit à petit du terrain, notamment chez Samsung, Global Foundries et NXP. On a aussi déjà vu le rôle des neurones à impulsion et des memristors pour améliorer l'équation énergétique du deep learning.

La mécanique et l'énergie sont les talons d'Achille non pas de l'IA qui est distribuable là où on le souhaite mais des robots. Un homme a une autonomie d'au moins une journée en état de marche convenable sans s'alimenter. Un robot en est encore loin.

D'où l'intérêt des travaux pour améliorer les batteries et notamment leur densité énergétique. Un besoin qui se fait sentir partout, des smartphones et laptops aux véhicules électriques en passant par les robots. Les progrès dans ce domaine ne sont pas du tout exponentiels. Cela a même plutôt tendance à stagner. Dans les batteries, c'est la loi de l'escargot qui s'appliquerait avec un quadruplement de la densité tous les 20 ans (**source**).

---

[659] Voir Energy and Policy Considerations for Deep Learning in NLP par Emma Strubell et al, juin 2019 (6 pages).

[660] Voir ORNL researchers design novel method for energy-efficient deep neural networks, mars 2018 et TensorQuant - A Simulation Toolbox for Deep Neural Network Quantization par Dominik Marek Loroch, 2017 (10 pages), un outil qui sert à benchmarker les méthodes de quantisation en fonction de la topologie des modèles de réseaux de neurones utilisés.

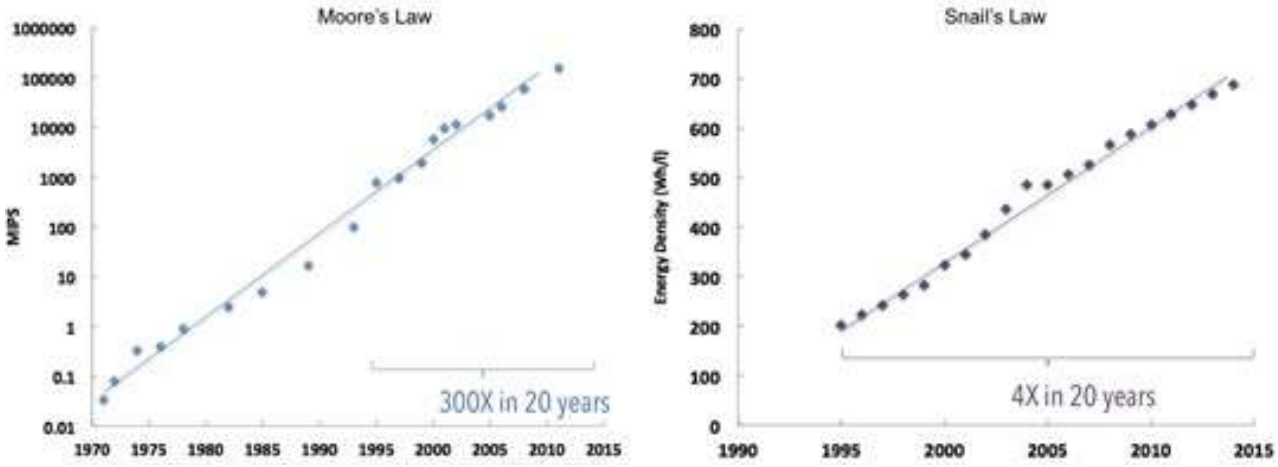
[661] Voir Green.ai par Roy Schwartz & Al, 2019 (12 pages).

[662] Voir Top AI Experts Create CodeCarbon, a Tool to Track and Reduce Computing's CO2 Emissions par BCG, décembre 2020.

Des laboratoires de recherche inventent régulièrement des technologies de batteries battant des records en densité énergétique ou du côté du temps de chargement, à base de matériaux différents et/ou de nano-matériaux, ou de composés différents au lithium.

Il y a notamment le lithium-sulfure ou le lithium-oxygène permettant en théorie d'atteindre une densité énergétique 20 fois supérieure à celle des batteries actuelles, utilisées dans les véhicules électriques[663].

Mais en elles sortent rarement, faute de pouvoir être industrialisées à un coût raisonnable ou de bien fonctionner dans la durée. Parfois, on arrive à une densité énergétique énorme, mais cela ne fonctionne que pour quelques cycles de charge/décharge. Trop injuste !



Résultat, pour le moment, la principale voie connue est celle de l'efficacité industrielle, choisie par Elon Musk dans la création de sa Gigafactory dans le Nevada, une usine à $5B qui exploite la technologie de batteries standards de Panasonic, qui a aussi mis $1B au pot pour le financement de l'usine.

Une usine qui est aussi proche d'une **mine de Lithium**, à Clayton Valley, l'un des composants clés des batteries et qui devait démarrer sa production en 2020. En novembre 2019, Elon Musk anonçait lancer une autre Gigafactory, près de Berlin en Allemagne. La construction devrait démarrer en 2020 pour une mise en production en 2021. On peut probablement ajouter une ou deux années à ce planning !

On peut cependant citer l'étonnante performance d'un laboratoire de l'université de Columbia qui a réussi à alimenter un composant CMOS avec de l'énergie provenant de l'ATP (adénosine triphosphate), la source d'énergie principale des cellules vivantes qui est générée par les nombreuses mitochondries qu'elles contiennent. Cela ouvre des portes vers la création de solutions hybrides biologiques et informatiques insoupçonnées jusqu'à présent. La véritable bionique en quelque sorte !

Performance énergétique des processeurs

La bataille de la performance énergétique des processeurs bat son plein[664]. On évalue cette capacité en comparant les TFLOPS/watt pour l'entraînement et les TOPS/watt pour les inférences. Mais pour comparer des choux avec des choux pour ces TOPS/watt, il faut savoir de quoi l'on parle : d'entiers 8 bits, 4 bits ou de BNN ?

---

[663] Voir Stockage de l'électricité: les batteries lithium-ion face au tout pétrole de Nicolas Hahn, décembre 2013.

[664] Voir Machine Learning Emissions Calculator qui permet d'évaluer la consommation d'énergie d'un processeur en fonction de son type de du nombred'heures de fonctionnement, indépendamment des algorithmes utilisés.

La performance énergétique des chipsets se situe sur une échelle comprenant deux ordres de grandeur, située entre 10 GigaOps/W à 10 TeraOps/W comme on peut l'observer dans le schéma suivant, les types d'opérations étant indiqués avec les signes utilisés. Il montre sans surprise à droite les processeurs pour data-centers et à gauche les processeurs de l'embarqué[665].

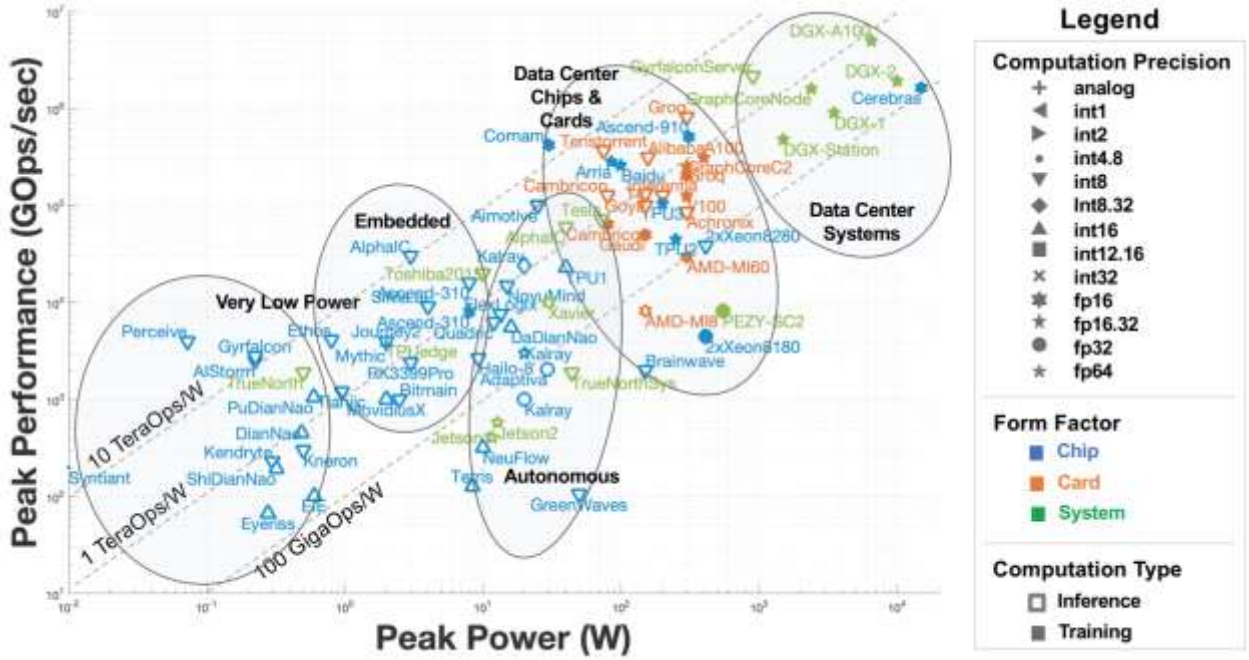
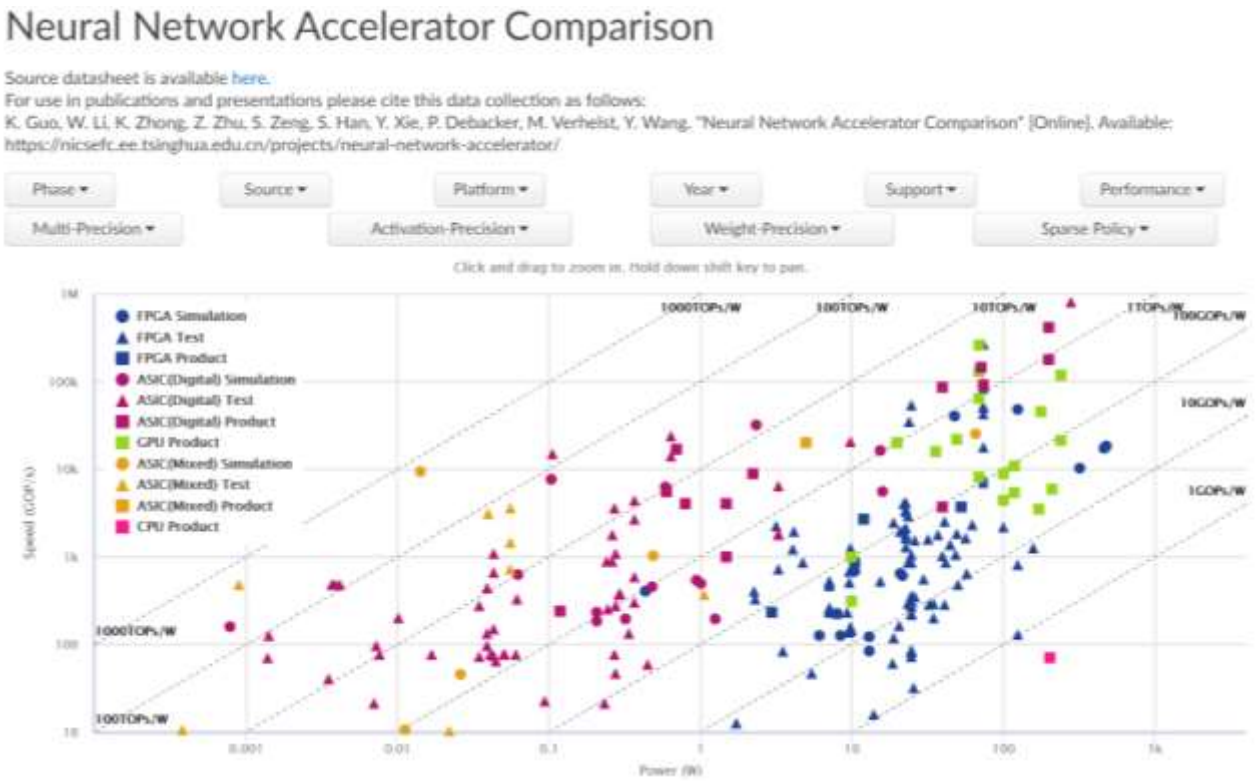


Fig. 2. Peak performance vs. power scatter plot of publicly announced AI accelerators and processors.

Dans cet autre chart, on constante sans surprise que l'efficacité énergétique des ASIC est meilleure que celle des FPGA, ce qui n'est pas une grande surprise pour les spécialistes[666].



[665] Voir Survey of Machine Learning Accelerators par Albert Reuther et al, septembre 2020 (11 pages).

[666] Voir Neural Network Accelerator Comparison pour le chart ci-dessous.

Approximate computing

L'approximate computing recouvre les techniques visant à réduire la précision des calculs du deep learning pour économiser énergie et également, améliorer la latence des inférences.

Elle fait partie du large champ des techniques d'optimisation des réseaux de neurones servant à en compresser la taille et réduire la complexité[667].

Elle s'appuie principalement sur le principe de la **quantisation** qui consiste à passer des nombres flottants aux nombres entiers au niveau des inférences voire de l'entraînement des modèles[668]. Cela permet au passage de réduire l'empreinte mémoire des inférences. De nombreux travaux de recherche ont été publiés dans ce sens, notamment pour la reconnaissance d'images avec des réseaux convolutifs[669]. Ils montrent en général une dégradation acceptable de la performance des modèles, mais c'est à évaluer au cas par cas et n'est pas valable dans toutes les situations[670].

Une méthode extrême consiste à passer carrément aux BNN, binary neuron network, qui utilisent des neurones binaires qui rappellent les perceptrons de 1957. C'est la simplification la plus extrême. C'est grâce à cette méthode que certains composants prototypes atteignent des records en TOPS/Watt[671]. Notons dans ce domaine que **Plumerai** (2017, UK) propose Larq, une bibliothèque Python de création de BNN. Elle est partenaire de son compatriote **XMOS** (2005, UK, $94,8M) qui développe des processeurs embarqués spécialisés pour les BNN dans les applications audio.

D'autres propositions consistent à optimiser le calcul en nombres flottants en utilisant de nouveaux format comme le BFLOAT16 proposé par Google qui augmente le nombre de bits utilisés pour l'exposant et moins pour la matisse. IBM propose de son côté l'usage de nombres flottants 8 bits hybrides qui pourrait apporter un nouveau gain de performance de x4. Reste à le supporter dans les chipsets [672]!

On peut aussi réduire le temps et la complexité des calculs avec le sparse computing qui optimise le traitement des matrices dites creuses en compressant les matrices avant le lancement des calculs. Nous avons vu que cela était notamment mis en œuvre dans le dernier chipset A100 de Nvidia. Les gains semblent assez appréciables.

énergie et optimisation

« approximate computing »

quantification : passer des flottants à des nombres entiers pendant ou après l'entraînement.

Binary Neuron Networks : neurones « binaires », un peu comme les Perceptrons.

sparse computing : éliminer les valeurs nulles.

in-memory computing : calculer directement dans la mémoire.

Domain Specific Architectures : architectures ad-hoc par usages.

edge AI : meilleure latence, optimisation des ressources


Green AI

Roy Schwartz[1] Jesse Dodge[1,2] Noah A. Smith[1,3] Oren Etzioni[1]

[1]Allen Institute for AI, Seattle, Washington, USA
[2]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[3]University of Washington, Seattle, Washington, USA

July 2019

Abstract

---

[667] De nombreuses techniques de simplification des réseaux de neurones sont couramment employées : l'élagage (pruning), la décomposition de tenseurs, les transformeurs, la distillation de connaissance, etc. Voir ce review paper An Survey of Neural Network Compression par James O'Neill, juin 2020 (73 pages).

[668] Voir Training and Inference with Integers in Deep Neural Networks par Shuang Wu et al, février 2018 (14 pages) ainsi que le review paper Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations par Itay Hubara et al, 2018 (30 pages).

[669] Voir And the Bit Goes Down: Revisiting the Quantization of Neural Networks par Pierre Stock et al, janvier 2020 (12 pages).

[670] Voir The Pitfall of Evaluating Performance on Emerging AI Accelerators par Zihan Jiang et al, 2019 (10 pages) qui évoque une perte de qualité des inférences avec une quantisation à 8 bits.

[671] Voir Imec and GLOBALFOUNDRIES Announce Breakthrough in AI Chip, Bringing Deep Neural Network Calculations to IoT Edge Devices, juillet 2020. Le composant proposé génère 2900 TOPS/W grâce à l'usage de BNN et d'optimisations matérielles diverses.

[672] Voir IBM Invests In AI Hardware par Karl Freund, mars 2020.

Côté matériel, nous l'avons aussi déjà évoqué, de nombreux travaux portent sur le rapprochement des unités de calcul de la mémoire, même si les gains sont actuellement difficiles à apprécier.

Refroidissement

Le refroidissement d'un data-center représente habituellement un coût énergétique non négligeable. Il représente plus de la moitié de l'énergie qui est engloutie par les serveurs et le stockage, avec facilement plus de 30% de la consommation électrique totale.

L'efficacité énergétique d'un data-center est souvent évaluée par son PUE (Power Usage Effectiveness). L'indicateur PUE a été lancé par le consortium **Green Grid** en 2007[673]. C'est le ratio entre la consommation totale d'énergie électrique d'un data center et celle qui alimente la partie informatique comprenant les serveurs, le stockage et la connectique réseau. Il devrait être idéalement de un, toute l'énergie allant aux ressources informatiques. En pratique, il est supérieur, ne serait-ce que pour alimenter les ventilateurs et/ou pompes des systèmes de refroidissement à air ou à eau, mais aussi les onduleurs.

$$\text{PUE} = \frac{consommation\ totale}{consommation\ IT} = 1 + \frac{consommation\ hors\ IT}{consommation\ IT} \qquad \text{DCiE} = \frac{consommation\ IT}{consommation\ totale}$$

On peut aussi utiliser le DCiE (data-center infrastructure efficiency) qui doit aussi être proche de 1, mais par le bas.

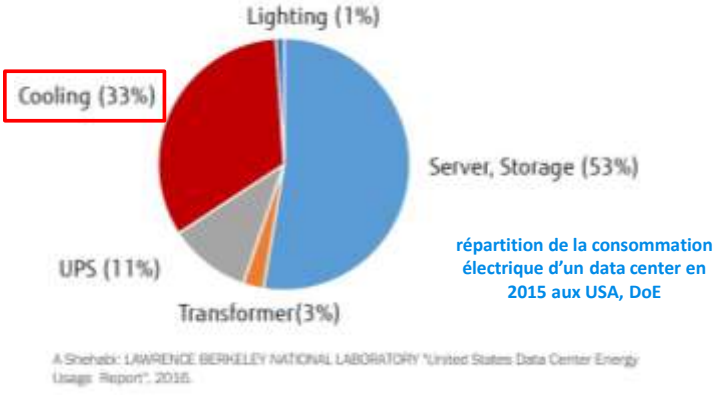
Chez **Google**, le PUE atteint 1,1 (courbe *ci-contre*) [674]. En Suède, **Facebook** arrive à descendre à un PUE de 1,05 grâce à des sources froides plus importantes.

On peut ajouter au PUE le ERE, qui indique l'Energy Reuse Effectiveness, la proportion d'énergie dégagée par les équipements informatiques qui peut être réutilisées à l'extérieur du data-center, par exemple comme énergie de chauffage.

$$\text{ERE} = \frac{énergie\ réutilisée}{consommation\ IT}$$

Jusqu'il y a quelques années, les racks de data-centers consommaient en moyenne moins de 10kW. Avec les processeurs dédiés à l'IA, la donne change.

Le chipset Nvidia A100 dégageant 400W, un serveur DGX A100 en dégage au moins huit fois ça, donc 3,2 kW rien que pour les GPU. Au total, ils consomment 6,5 kW. Pour un rack en comprenant quatre, il va donc falloir dégager quasiment 26 kW en chaleur.



répartition de la consommation électrique d'un data center en 2015 aux USA, DoE

Mais on a ici 20 PFLOPS/s (en flottants 16 bits) au compteur à ce compte-là !

[673] Voir Green Grid Data Center Power Efficiency Metrics: PUE and DCiE, 2007 (16 pages).

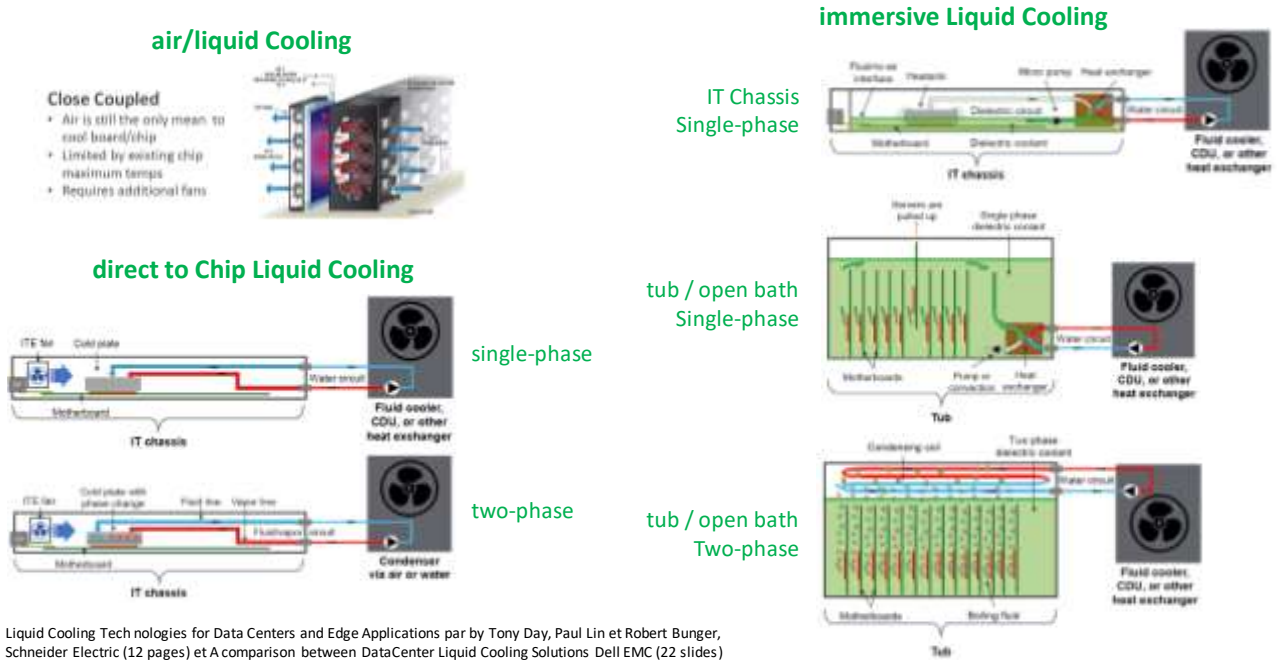
[674] Source : https://www.google.com/about/datacenters/efficiency/.

Il existe deux techniques principales de refroidissement : par air et par eau. Les serveurs Nvidia DGX sont refroidis par air, avec une architecture classique pour les racks comprenant des ventilateurs à larges pales en façades, de grands dissipateurs de chaleur placés sur les modules SMX (les cartes électroniques comprenant le processeur ; on voit quatre de ces dissipateurs dans la photo *ci-contre*), le flux d'air traversant tout le rack. C'est l'approche utilisée dans le supercalculateur Joliot Curie en France qui est équipé de V100.

Certains partenaires de Nvidia ont cependant conçu leurs propres serveurs refroidis par eau comme **Baidu** qui a créé un bloc de refroidissement par eau pour les modules SMX[675]. Autre exception, la station de travail DGX Station équipée de V100 qui est aussi refroidie par eau. Les cartes graphiques Nvidia de la série GeForce RTX peuvent aussi être refroidies par eau. Elles sont notamment utilisées dans des serveurs de cloud gaming ou pour réaliser des inférences d'IA.

Le refroidissement par eau est de plus en plus utilisé dans les data centers, notamment en raison de l'augmentation de la puissance thermique à dégager par rack. C'est lié au fait que l'eau peut transporter bien plus de calories que l'air[676]. Par contre, il faut gérer différents défis : éviter les fuites, la corrosion, la croissance microbiologique et les dépôts de tartre. Cela passe par l'usage d'eau filtrée, déminéralisée et déionisée, et par l'usage de tuyauteries spéciales, notamment en alliages de cuivre ou en aciers inoxydables.

Les techniques de refroidissement par liquides sont variées comme illustré dans les schémas *ci-dessous*[677]. Le refroidissement air/eau exploite de l'air pour refroidir l'électronique et de l'eau pour refroidir l'air avec un échangeur de chaleur intégré dans le rack. Le refroidissement direct du chipset par eau ou un autre liquide caloporteur comme le propylène glycol, souvent utilisé pour refroidir des cartes graphiques PCI.



Liquid Cooling Tech nologies for Data Centers and Edge Applications par by Tony Day, Paul Lin et Robert Bunger, Schneider Electric (12 pages) et A comparison between DataCenter Liquid Cooling Solutions Dell EMC (22 slides)

---

[675] Voir Baidu X-MAN Liquid Cooled 8-Way NVIDIA Tesla V100 Shelf par servethehome, 2018.

[676] Voir New Horizons: Dale Sartor, P.E. Getting (Back) to Liquid Cooling par Dale Sartor et al, octobre 2019 (51 slides) et Not Just About Chip Density – 5 Reasons to Consider Liquid Cooling for Your Data Center par Wendy Torell, Schneider Electric, juillet 2019.

[677] Source des dessins : Liquid Cooling Technologies for Data Centers and Edge Applications par Tony Day, Paul Lin et Robert Bunger, Schneider Electric (12 pages) et Voir A comparison between DataCenter Liquid Cooling Solutions par Paolo Bianco, 2019 (22 slides).

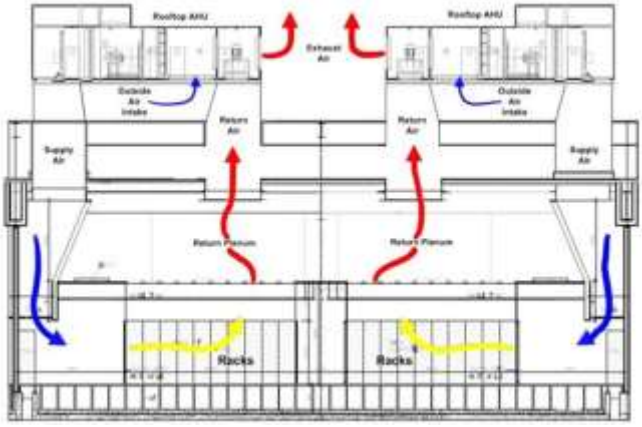
Puis existent des méthodes de refroidissement par immersion à une ou deux phases, avec un échange de chaleur passant par le liquide ou via un condensateur refroidissant la phase gazeuse du liquide.

Des techniques de refroidissement par eau chaude exploitent des cycles thermiques plus efficaces allant de 32°C (eau en arrivée) à 42°C (eau en sortie de racks), comme ce qui est mis en œuvre dans le supercalculateur Jean Zay du GENCI pour 90% de ses équipements.

L'eau chaude générée par les racks peut être exploitée pour chauffer ensuite le bâtiment qui l'accueille et des bâtiments voisins[678].

Le **free cooling** est une technique de refroidissement de plus en plus courante dans les data-centers depuis une dizaine d'années. Au lieu d'avoir un cycle fermé d'air qui est refroidit par des échangeurs de chaleur, de l'air frais est récupéré à l'extérieur du bâtiment et l'air réchauffé y est également dégagé[679]. Par contre, le data-center doit avoir été architecturé pour[680].

Cela fonctionne dans les latitudes où le climat est frais ou tempéré avec un air extérieur de température inférieure à 30°C. La charge électrique est ainsi concentrée sur les ventilateurs, un peu comme la solution adiabatique mise en œuvre chez Scaleway. C'est la technique utilisée par **OVH** pour son récent centre de calcul RBX8 ouvert à Roubaix et annoncé en octobre 2020. Les serveurs y sont refroidis par eau et l'eau l'est par l'air extérieur. Cela évite les désagréments de l'humidité de l'air qui pourrait endommager les serveurs.



Il existe d'autres méthodes d'échange air/air pour isoler l'air extérieur de l'air intérieur. C'est ce que l'on appelle le free cooling indirect.

L'opérateur de cloud français **Scaleway** (filiale de Free) a mis en place dans son data center DC5 situé en région parisienne une variante du free cooling qui exploite de l'air extérieur, qui est réfrigéré par évaporation d'eau de manière adiabatique en été. Cela fonctionne même lorsqu'il fait plus de 30°C dehors. En hiver, l'air extérieur doit être amené à plus de 16°C, ce qui est obtenu en le mélangeant avec l'air sortant des racks. Le centre était partiellement équipé (au 1/11e) lorsque j'ai pu le visiter en juillet 2020, avec 1,8 MW de puissance machine installée pour un objectif final de 22 MW. Les défis sont de maîtriser la qualité de l'eau pour éviter les phénomènes de corrosion. Ils la produisent eux-mêmes avec un adoucisseur à permutation ionique, de la désinfection contre les micro-organismes, du filtrage de sédiments, de la déchloration et une osmose inverse. Ils en consomment 15 000 m^3 par an et disposent d'un réservoir de 234 m^3. Le système utilise des filtres en cellulose d'origine Camfil, qui doivent être remplacés tous les 6 mois et qui débarrasse l'air extérieur de ses poussières et particules fines. Le système permet d'économiser au moins trois quart de l'électricité qui serait nécessaire à un refroidissement classique[681]. Le PUE du DC5 est inférieur à 1,15.

---

[678] Voir Temperature Management in Data Centers: Why Some (Might) Like It Hot par Nosayba El-Sayed et al (12 pages) qui explique les tenants et aboutissants du refroidissement à haute température et Cooling Control Strategies in Data Centers for Energy Efficiency and Heat Recovery par Riccardo Lucchese, 2019 (76 pages).

[679] Source de l'illustration : Free Cooling: the Server Side of the Story par Johan De Gelas, février 2014.

[680] Voir The hidden cost of free-cooling and what you can do, ABB (12 pages). Il évoque la consommation d'énergie des ventilateurs des serveurs qui croît au carré de leur vitesse et les fuites (« leakage ») des processeurs qui sont plus élevées à haute température.

[681] Voir cet excellent compte-rendu ultra-détaillé de visite : Visite du data center Scaleway DC5 (refroidissement adiabatique) par Vivien, janvier 2019.

Autre solution, testée en Ecosse par **Microsoft** depuis 2018 en collaboration avec le groupe Français **Naval Group**, installer les data-centers sous la mer et profiter de son réservoir presque infini d'eau froide[682]. Les containers cylindriques étaient remplis d'azote qui est moins corrosif que l'oxygène de l'air ce qui a permis de rendre les serveurs plus fiables avec un taux de panne de 1/8e des taux de panne classiques. Pour que le système soit opérationnel dans la durée, il faudrait probablement que ce taux de panne soit encore plus bas.

Cybersécurité

La cybersécurité est aussi un grand terrain de jeu pour l'IA. Cette dernière intervient à trois niveaux :

- L'IA est vulnérable à de nouvelles cyberattaques qui exploitent le fonctionnement interne des réseaux de neurones[683].
- L'IA permet de créer de nouvelles attaques affectant les réseaux et systèmes d'information.
- L'IA permet de prévenir et détecter les attaques. Le machine learning peut détecter les menaces qu'il s'agisse de spams d'emails, de mails de phishing ou d'identification de vulnérabilités diverses dans les réseaux et systèmes d'information, notamment les « non malware attacks » (remote login, attaques par scripts et macros, etc.).

Protections à base d'IA

C'est encore du côté des startups que nous allons faire le tour des outils de la cybersécurité exploitant peu ou prou de l'IA[684].

Les tentatives de phishing sont détectées par **GreatHorn** (2015, USA, $21,8M) ou avec **Lookout** (2007, USA, $282M) qui sécurise les mobiles avec un modèle prédictif. Les malwares sont détectés avec du machine learning par **Cylance** (2012, USA, $297M), une startup financée entre autres par InQTel, le fonds de corporate venture de la CIA. La startup a été acquise par Blackberry en novembre 2018.

Des chercheurs de **Google** démontraient début 2020 une solution à base de machine learning capable de mieux détecter des documents malicieux attachés à des mails. Ceux-ci sont majoritairement des fichiers Office (surtout Word et Excel). Le nouveau scanner détecterait 10% de documents Office en plus[685].

DeepInstinct (2014, USA/Israël, $37M) protège les systèmes contre les failles de sécurité récentes ("zero day threats"). Ce serait la première startup à exploiter le deep learning - avec des GPU Nvidia - tandis que la plupart utilisaient du machine learning jusqu'à présent pour faire de l'analyse multifactorielle des menaces en lieu et place de l'utilisation de bases de signatures de virus. **Interset** (2015, Canada, $24M) est dans le même créneau.

---

[682] Voir Microsoft finds underwater datacenters are reliable, practical and use energy sustainably par John Roach, septembre 2020.

[683] Voir cette excellente présentation OK robot, machine learning and cybersecurity, Alexander Polyakov, ERPScan, 2018 (60 slides) qui décrit bien les menaces liées à l'IA et la manière dont l'IA améliore aussi la cybersécurité. Et aussi dans cette vidéo par un spécialiste de la cybersécurité de SAP.

[684] Voir IA et cybersécurité : info ou intox ? - ZDNet par Louis Adam, novembre 2018.

[685] Voir Gmail Is Catching More Malicious Attachments With Deep Learning par Lily Hay Newman, février 2020.

SentinelOne (2013, Israël, $696M) utilise le machine learning pour détecter et supprimer les menaces connues et inconnues (dites zero-day) affectant les terminaux fixes et mobiles et pour toutes les méthodes d'attaques (par fichiers, injection mémoire, scripting...). Avec l'agent Singularity, ils couvrent ainsi l'EPP (Endpoint Protection Platform) pour protéger les postes de travail et smartphones et l'EDR (Endpoint Detection and Response) qui permet de parer aux attaques. Le tout exploite l'analyse comportementale des vecteurs d'attaques suspects. La solution entend remplacer les antivirus traditionnels. C'est notamment un concurrent de Cylance.

DarkTrace (2013, UK, $179,5M) utilise le machine learning pour détecter les menaces d'intrusion dans le système d'information de l'entreprise. Ils exploitent notamment du machine learning non supervisé pour détecter les anomalies de comportement à tous niveaux, dans les serveurs, desktop/laptops, et réseaux, y compris les "zero day threats", ces attaques lancées immédiatement après la publication de failles dans les systèmes d'exploitation et infrastructures.

Ils neutralisent ensuite les menaces avec leur "antigène" maison. L'ensemble s'interface avec les systèmes industriels au standard SCADA.

Dans le même genre, **Recorded Future** (2009, USA, $57,9M) utilise le machine learning pour détecter les menaces de sécurité en temps réel. La solution Deep Armor de **SparkCognition** (2013, USA, $56,3M) utilise aussi le deep learning et pour protéger les objets connectés dans l'industrie et les applications critiques. **CrowdStrike** (2011, USA, $481M) est une autre solution (en cloud) de protection des infrastructures d'une entreprise, détectant les attaques. **Ogo Security** (France) est aussi sur ce créneau, protégeant les ressources informatiques des PME.

Des startups comme **Onfido** (2012, UK, $188M) vérifient l'identité de clients de service en ligne. C'est de la détection de fraude basée sur du machine learning et du prédictif.

Shape Security (2011, USA, $183M) propose deux produits : Defense, qui détecte les attaques et fraudes contre les sites marchands et applications mobiles (fraudes aux cartes cadeaux, vol de contenu, DDoS) et Blackfish, qui évite l'usage de mots de passe volés et les attaques par force brute.

Fortscale (2014, Israël, $23M, acquis par RSA Security en 2018) identifie de son côté les menaces internes dans les entreprises, avec sa solution User & Entity Behavioral Analytics (UEBA). Il va détecter des comportements suspects comme la copie de fichiers de grande taille sur des clés USB ! Dans les pays où ce genre de surveillance est autorisé !

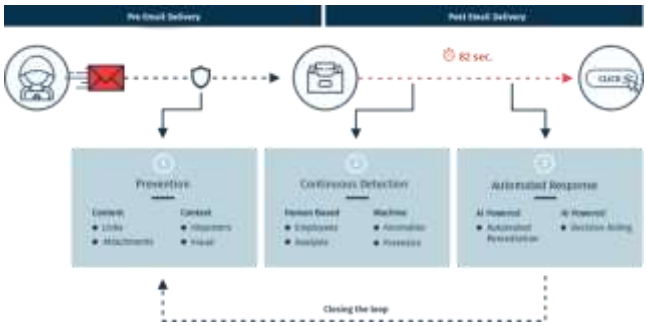
Des startups se sont aussi spécialisées sur la cybersécurité des objets connectés. C'est le cas de l'américain **SparkCognition** (2013, USA, $163,69M) dont l'offre s'articule autour de DeepArmor, une solution d'antivirus qui ne s'appuie pas sur un dictionnaire de signatures à mettre à jour régulièrement. Elle est complétée par SparkPredict qui teste de nombreux paramètres et variables de fonctionnement de systèmes embarqués pour détecter leurs failles de sécurité.

Beyond Security (1999, Israël) propose une solution dans le même registre qui teste tous les effets de bord de protocoles réseaux et logiciels pour identifier des trous dans la passoire des objets connectés. Les opérateurs télécoms sont aussi intéressés par d'autres formes de fraudes. Ainsi, Orange utilise la solution **Skymind** (2014, USA, $6,3M) pour détecter la fraude aux cartes SIM en exploitant les logs d'appels via un réseau de neurones utilisant un autoencodeur.

CybelAngel (2013, France, 51,7M€) a développé une solution de screening en continu du dark web sur les menaces pesant sur les entreprises clientes. Elle détecte les fuites de données de l'entreprise, qu'elles aient été volées ou aient circulé par inadvertance. La solution scanne le deep et le dark web ainsi que les sources de stockage de données connectées. Elle associe du machine learning et de l'expertise humaine. La société aurait déjà Sanofi, Louis Vuitton et L'Oréal comme clients.

DrawBridge (2010, USA, $68,7M) a créé DrawBridge Identity Platform, un système d'identification des utilisateurs multi-devices à base de machine learning et connecté aux sources d'informations de l'entreprise.

IronScales (2013, Israël, $31M) utilise le machine learning pour détecter les tentatives d'intrusion par phishing. La solution est compatible avec Office 365 de Microsoft et la G-suite de Google et s'y installe sous la forme d'extensions, en plus des détections réalisées côté serveur et cloud. Leur solution détecte même les attaques utilisant du « social engineering ».

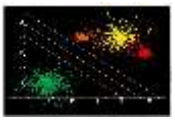


CloudConstable (Canada) propose une solution de protection à base d'IA pour les utilisateurs vulnérables, notamment les enfants et les personnes âgées. Le tout avec un service en cloud sur abonnement.

En 2019, **Microsoft** lançait Azure Sentinel, une solution de supervision de sécurité dotée de modules de machine learning en cloud qui s'interface avec les solutions de Cisco, Check Point et Symantec[686].

IBM qui met Watson à toutes les sauces l'a aussi décliné dans la cybersécurité. Leur QRadar Advisor with Watson analyse toute la littérature disponible sur la cybersécurité pour aider les entreprises à détecter et circonscrire les menaces. La solution détecte les attaques dans les nœuds de réseaux, analyse les incidents et gère les réponses aux détections d'incidents de sécurité.

Cisco propose de son côté Encrypted Traffic Analytics, une solution logicielle qui identifie des menaces dans des données cryptées via la détection d'anomalies dans les métadonnées. C'est complété par leur outil d'analytics Stealthwatch. Cisco a fait l'acquisition en juin 2019 de **Sentryo** (2014, France, 12M€), spécialisée dans la sécurité de réseaux industriels compatible avec SCADA.

	DARKTRACE	onfido	CISCO
détection de malware à base de machine learning **startup financée par InQTel (CIA)** 2002, USA, $297M, acquise par Blackberry en novembre 2018	**détection de menace par identification des comportements anormaux** **compatibles avec systèmes industriels (SCADA, …)** 2013, UK, $230M	**vérification d'identité multi-critères : pièces d'identité, reconnaissance du visage, géolocalisation** **startup de San Francisco** 2012, UK, $110M	**Encrypted Traffic Analytics : détecte des menaces dans des données cryptées via la détection d'anomalies dans les métadonnées.** **Stealthwatch : analytics**
			

Les fonctions de machine learning de **Check Point** permettaient de détecter un agent malveillant « Agent Smith » en 2020 qui ciblait Android avant qu'il ne devienne dangereux et soit déployé. Leur solution ThreatCloud AI identifie des corrélations entre des événements suspects et malveillants issus de plusieurs sources pour détecter des infections et identifier les machines infectées. Il détecte aussi les événements atypiques qui requièrent une attention et des actions correctives rapides.

Citons une approche originale de l'IA, la stylométrie qui permet d'identifier les développeurs de code en général, et de code malicieux en particulier ! C'est une nouvelle arme contre les pirates qui est pour l'instant encore au stade de la recherche[687].

---

[686] Voir Azure Sentinel, un SIEM augmenté à l'IA signé Microsoft - Le Monde Informatique par Dominique Filippone, mars 2019.

[687] Voir De-anonymizing Programmers via Code Stylometry, avec comme contributrice Aylin Caliskan-Islam, août 2015 (17 pages) et Grâce à l'apprentissage automatique, le style des programmeurs est facilement reconnaissable, août 2018.

L'algorithme aurait une précision de 83% ce qui est un peu faible. Il permettrait, en plus de la capacité à identifier les auteurs de virus et malware, de lutter également contre le plagiat.

Dathena Science (2016, Singapour) est une startup créé par des Français qui détecte les stocks et flux de données confidentielles dans les entreprises grâce à des techniques de traitement automatisé du langage (NLP) et à du machine learning permettant de détecter des anomalies de flux.

Ubble (France) vérifie les identités des consommateurs en ligne avec de l'IA et de la vision artificielle, pour de la détection de visages et de l'analyse de documents officiels.

Datadome (2015, France, 3,6M€) propose une solution de protection temps-réel contre les bots qui attaquent les sites web. Le système fonctionne en SaaS avec les sites hébergés un peu partout (AWS, …). Il détecte les bots en temps réel et leur type de comportement : scrapping de site web, déni de service, attaque sécurité par force brute ou détection de vulnérabilité. Le tout est évidemment complété d'abondants outils d'analytics.

On peut enfin intégrer dans ce long panorama les solutions de détection de personnes à base de deep learning qui sont intégrés dans les caméras de surveillance ou dans leurs logiciels, comme chez **D-Link**.

Vulnérabilités de l'IA

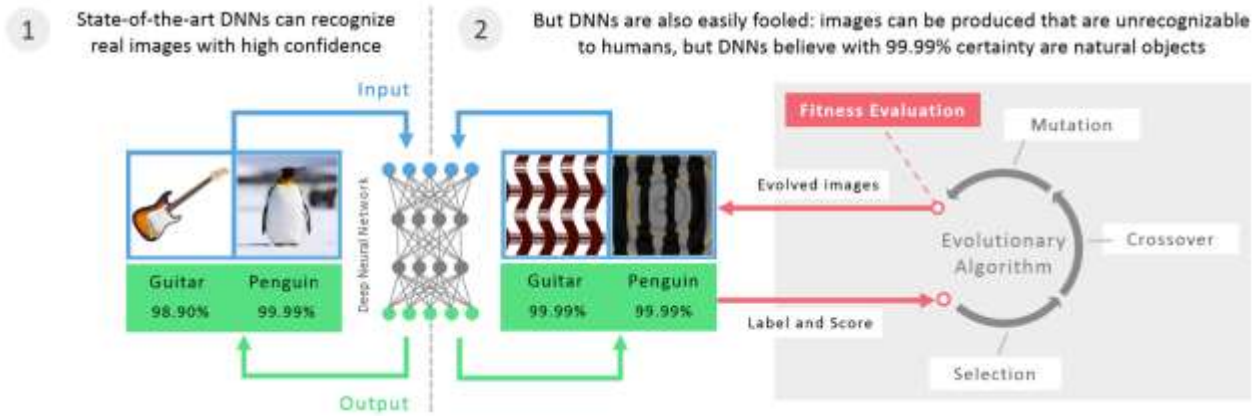
L'intelligence artificielle permet aussi de créer de nouvelles menaces. Les algorithmes de machine learning et de deep learning peuvent être retournés contre eux-mêmes par des pirates, en étant alimenté par des données bidouillées qui altèrent leurs sens. Ces attaques peuvent intervenir tout d'abord au niveau des capteurs ou des réseaux pour injecter des données modifiées.

Les vulnérabilités de l'IA interviennent à tous les étages de sa chaîne de valeur : dans les données qui l'alimentent, dans les capteurs qui les ont générées, dans les télécommunications qui les ont transmises et dans les modèles qui les ont traitées.

Les réseaux de neurones de vision artificielle peuvent être trompés avec des images modifiées par une technique à base de deep learning voisine de la stéganographie, qui n'en change pas l'apparence pour la vision humaine (*exemple ci-contre*)[688]. Ainsi, dans cet exemple très connu, un panda légèrement modifié devient-il un singe pour l'algorithme de deep learning[689].

Cela vient des méthodes de réseaux de neurones convolutifs et de leurs techniques de représentation hiérarchiques qui ne correspondent pas à la décomposition fonctionnelle humaine mais à des méthodes que l'on pourrait qualifier de « plus mathématiques » et qui sont contournables.

L'entraînement d'un réseau de neurones convolutif consiste à minimiser les erreurs de classification en modifiant les poids des neurones. Les attaques adversariales consistent à faire l'inverse : elles modifient les images pour augmenter le taux d'erreur de la reconnaissance[690].



Le système de login de l'iPhone X par reconnaissance 3D du visage pouvait lui aussi être trompé par une fausse tête en plastique imprimée en 3D où étaient collées des images du nez, de la bouche et des yeux de l'utilisateur, dans le genre Mission Impossible. Le tout, pour $150 ! Ce hack provenant de la société de cybersécurité vietnamienne **Bkav**[691] avait été réalisé quelques jours après la mise sur le marché du nouvel iPhone X fin 2017 !

---

[688] Voir Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples, février 2016. Et un exemple plus récent avec Query-efficient Black-box Adversarial Examples, de Andrew Ilyas, Logan Engstrom, Anish Athalye et Jessy Lin, avril 2018 (8 pages) qui présente une technique équivalente mais très rapide. Et puis aussi High dimensional spaces, deep learning and adversarial examples de Simant Dube, avril 2018 (29 pages), qui décrit les modèles mathématiques de perturbation d'images par réseaux adversariaux. Puis Adversarial Patch, mai 2018 (6 pages) qui décrit une méthode d'attaque ciblée de parties d'images. Et enfin, Security and Privacy in Machine Learning par Nicolas Papernot, décembre 2017 (99 slides).

[689] Voir Council Post: Are You Ready For The Age Of Adversarial AI? Attackers Can Leverage Artificial Intelligence Too par Rahul Kashyap, janvier 2020.

[690] Voir Explaining and Harnessing Adversarial Examples, Ian Goodfellow et al, 2015 (11 pages). Voir aussi Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, Nguyen et al, 2015 (20 pages).

[691] Voir Bkav's new mask beats Face ID in "twin way": Severity level raised, do not use Face ID in business transactions, novembre 2017 et la vidéo associée.

Depuis, Apple a perfectionné son Face ID pour éviter ce genre d'attaque. Les méthodes anti-spoofing peuvent par exemple vérifier que le visage de la personne est bien vivant et animé[692]. Par ailleurs, FaceTime ne détecte pas la version masquée des utilisateurs ce qui est gênant pendant les périodes covidesques.

Histoire d'être équitable, juste après cet exploit, c'était au tour du face login de Windows 10 d'être hacké par une simple photo[693] ! À l'envers, d'autres chercheurs ont réussi à modifier des images de visages pour qu'elles ne soient pas reconnues par des IA, mais toujours reconnues par l'œil humain[694]. Dans un autre domaine, le robot **Pepper** de Softbank Robotics serait aussi très vulnérable aux attaques de pirates du fait de nombre de failles de sécurité au niveau réseau[695]. Ces nouvelles menaces liées à des détournements de l'IA vont faire émerger à leur tour de nouvelles parades.

Entre temps, des chercheurs ont aussi montré qu'il était facile de récupérer les empreintes digitales d'un utiliser pour pouvoir ensuite forcer le login sur son smartphone. Sauf dans le cas où la lecture d'empreintes digitales passe par la détection des veines sous-cutanées, ce qui n'est pas courant.

Et les exemples ne manquent pas de la manière dont on peut tromper les systèmes de reconnaissance d'images à base de réseaux de neurones [696]! Le cas le plus édifiant est celui de cette entrepreneuse chinoise, Dong Mingzhu, qui se voyait pénalisée par une contredanse dans la ville de Ningbo en 2018 car elle avait traversé une avenue hors des clous. C'était en fait son image qui était sur une publicité sur un bus [697]! Le réseau de neurones n'avait intégré le contexte !

Un réseau de neurones génératif peut aussi créer des empreintes digitales factices qui vont tromper certains systèmes de sécurité[698]. L'exemple de **DeepMasterPrints** issu de l'Université de New York arrive à tromper un système de reconnaissance d'empreintes digitales dans trois quarts des cas.

Autre cas bien connu, l'ajout de stickers sur des panneaux de signalisation qui peut tromper les systèmes de reconnaissance d'image de véhicules semi-autonomes ou semi-autonomes, comme l'a démontré McAfee avec le système Mobileye des Tesla[699].

---

[692] Voir About Face ID advanced technology, Apple.

[693] Voir Windows Hello Facial Recognition Bypassed with a Photo de Ryan Whitwam, décembre 2017. Mais la startup **TrueFace.ai** semble avoir trouvé une parade, aussi à base de deep learning.

[694] Voir Natural and Effective Obfuscation by Head Inpainting, 2018 (16 pages).

[695] Voir Le robot Pepper, nid à vulnérabilités de sécurité de Dominique Filippone, mai 2018 et la source de l'article : Adding Salt to Pepper - A Structured Security Assessment over a Humanoid Robot, 2018 (8 pages).

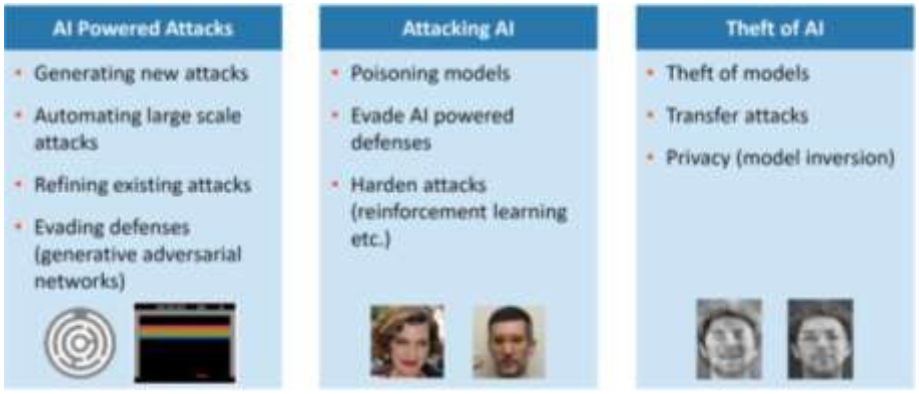
[696] Voir Des chercheurs ont mis en avant les faiblesses de l'IA en lui proposant des images confuses par Yvonne Gangloff, 2019.

[697] Voir AI Mistakes Bus-Side Ad for Famous CEO, Charges Her With Jaywalking par Tang Ziyi, novmebre 2018.

[698] Voir AI can create synthetic fingerprints that fool biometric scanners, Kris Holt, novembre 2018.

[699] Voir Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles par Steve Povolny et Shivangee Trivedi, McAfee, février 2020.

Une autre menace concerne le vol d'IA qui peut toucher celui de modèles de deep learning déjà entraînés[700] ou le croisement de modèles prédictifs et de bases de données qui permettent de désanonymiser des données anonymes. C'est peu plausible mais pas entièrement exclu.

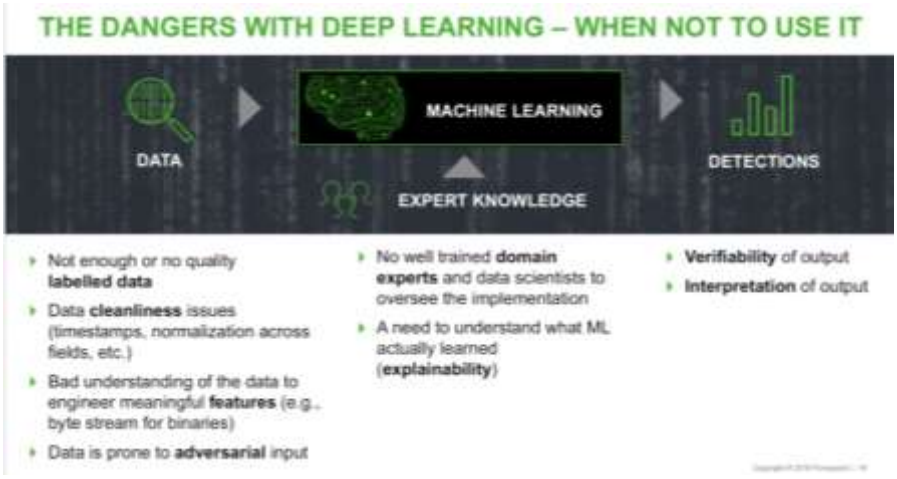


Il existe même des contre-indications à l'usage du deep learning comme présenté *ci-dessous*[701] : l'insuffisance de données propres et labellisées, la mauvaise compréhension de leur structure et dispersion, le manque d'experts dans le domaine, la difficulté à expliquer les algorithmes utilisés et à vérifier la validité des résultats.

Du côté des contre-mesures, des chercheurs ont par exemple trouvé le moyen d'encoder des images JPEG pour éviter leur modification destinée à tromper les réseaux convolutifs.

La **Blockchain** pourrait aussi être utilisée pour garantir la chaîne de confiance de transport de l'information alimentant les IA. Bref, c'est comme dans l'armement.

Des mesures de défense amènent à la création des contre-mesures et de leurs propres contre-mesures dans une course sans fin !



Dans la cybersécurité, la tranquillité absolue est une vue de l'esprit [702] !

Attaques à base d'IA

L'IA est aussi un nouvel outil pour les cyber-attaquants[703].

---

[700] Voir AI and Cybersecurity par Sridhar Muppidi et Koos Lodewijkx, IBM, 2018 (27 slides) d'où vient le slide.

[701] Le slide est issu de la présentation AI &ML in cybersecurity, why algorithms are dangerous par Raffael Marty Forcepoint, 2018 (41 slides). Voir aussi Security and Privacy Issues in Deep Learning, par Ho Bae et al, décembre 2018 (35 pages).

[702] Tout cela est bien documenté dans le rapport The Malicious Use of Artificial Intelligence - Forecasting, Prevention, and Mitigation qui met en avant les risques sécuritaires générés par l'IA et les moyens de commencer à les prévenir. A ce sujet, voir Encore une fois, les commentaires sur un rapport sur les IA passent à côté de l'essentiel, en privilégiant l'angle de l'effroi de Aymeric Poulain Maubant qui commente les réactions effarouchées et anxyongènes à ce rapport, février 2018.

[703] Voir Understanding the Strategic and Technical Significance of Technology for Security Implications of AI and Machine Learning for Cybersecurity, The Hague Security Delta, qui est une sorte d'équivalent du CLUSIF aux Pays-Bas, 2019 (40 pages) et Can Artificial Intelligence Power Future Malware?, par O. Kubovič & Al, 2018 (16 pages).

L'IA permettrait à des attaquants de mieux se protéger contre les recherches et les outils de défense en détectant les intrus.

Elle pourrait servir à diffuser des contenus ciblés et personnalisés avec des réseaux génératifs de type GANs comme des emails de phishing exploitant les sujets d'intérêt de la cible identifiés dans les réseaux sociaux, à créer de faux profils pour les réseaux sociaux avec des visages générés par des GANs impossibles à retrouver avec des moteurs de recherche.

L'IA peut aussi servir à créer des malwares encore plus puissants et difficiles à détecter par les systèmes de cybersécurité du marché[704]. Ces malwares peuvent imiter des patterns ressemblant au trafic normal chez la cible.

L'IA est aussi à la base de tromperies des humains. Nous traiterons à part du cas des fake news. Les contenus fictifs peuvent aussi servir à générer des arnaques redoutables. C'est le cas de l'arnaque au Président qui exploite les progrès des systèmes de génération de voix synthétique à base de GANs entraînés par la voix du président à partir de ses prises de parole publique[705].

[704] Voir DeepLocker: How AI Can Power a Stealthy New Breed of Malware par Marc Stoecklin, 2018.

[705] Voir Deepfake : l'arnaque boostée aux réseaux neuronaux par Louis Adam, septembre 2019.

Applications génériques de l'IA

Après avoir décrit les techniques de base de l'intelligence artificielle côté logiciels, algorithmes, données et infrastructures, nous allons passer aux applications et aux usages.

Cette première partie est dédiée aux applications génériques de l'IA qui sont généralement transversales aux entreprises de tous secteurs d'activité.

Elles sont organisées en deux grandes catégories :

- La **vision**, le **langage** et la **robotique** qui s'appuient sur les briques fondamentales de l'IA vues précédemment.
- Le **marketing**, la **vente**, les **ressources humaines**, la **finance et les achats** qui sont des fonctions horizontales dans les entreprises et font appel aux briques technologiques de l'IA ainsi qu'aux trois domaines précédents, en fonction des besoins.

Vision

La vision artificielle est l'application la plus courante et diversifiée de l'IA. C'est l'une des principales applications du deep learning. C'est dans ce domaine que les progrès de l'IA ont été les plus manifestes depuis 2012.

La recherche continue d'avancer dans ce créneau, en particulier dans les techniques de reconnaissance d'images pour élever au maximum le niveau sémantique de l'identification des personnes, objets et même terrains dans le cadre de la télédétection. On peut considérer que c'est le domaine le plus mature de l'IA.

Nous allons passer en revue les principaux usages de la vision artificielle et leurs progrès les plus récents.

Reconnaissance d'images

On la trouve pour les moteurs de recherche, les réseaux sociaux et les systèmes de sécurité et/ou vidéosurveillance. Elle est aussi utilisée couramment dans les appareils photos pour la mise au point et pour la détection des sourires.

État de l'art

La reconnaissance et la labellisation d'images sont des technologies qui progressent rapidement depuis 2012 grâce au deep learning et aux GPU.

Dans ses premières prouesses en 2012, le deep learning était capable de reconnaître un objet unique dans une image. L'état de l'art correspondant figure *à droite*[706]. Le réseau de neurones était entraîné sur une base de 1,2 millions d'images. Il pouvait reconnaître un millier de types d'objets différents mais avec un taux d'erreur de 37%, inacceptable pour tout système en production. Le réseau convolutif comprenait cinq couches et 60 millions de paramètres répartis dans 650 000 neurones. Il était entraîné sur des GPU Nvidia GTX 580 configurés avec 3 Go de mémoire.

Les images en entrée faisaient 224 pixels de côté et étaient en couleur. La base d'entraînement était issue de la base de référence d'images **ImageNet**. En 2016, cette dernière comportait plus de 10 millions d'images et atteignait 14,2 millions d'images en octobre 2019[707]. Elle sert à la réalisation de benchmarks de solutions de reconnaissance d'images.

Google et Facebook disposent de bases d'entraînement encore plus grandes, de plus de 100 millions d'images pour le premier et plus de 20 000 classes d'objets différentes. Les bases d'entraînement doivent régulièrement grandir en taille, même si Google et Facebook ne communiquent pas régulièrement dessus.

Trois ans plus tard, en 2015, il devenait possible d'identifier plusieurs objets et personnes dans une même image comme dans **DenseCap**[708] qui associe un réseau convolutif pour la détection des objets puis un autre pour leur labellisation. Il alimente un réseau à mémoire **LSTM** qui génère un label en texte clair des objets détectés.

Les images traitées faisaient 720 × 600 pixels.

---

[706] Il est issu de ImageNet Classification with Deep Convolutional Neural Networks de Alex Krizhevsky, Ilya Sutskever et Geoff Hinton de l'Université de Toronto (9 pages)

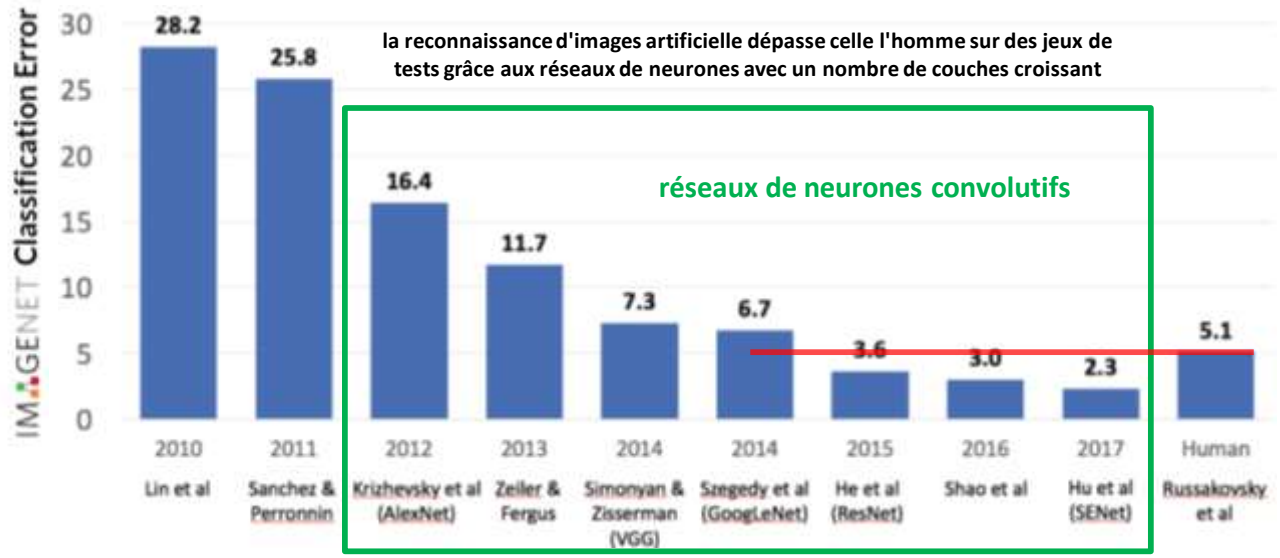
[707] L'augmentation de la base de référence ne change rien au dimensionnement du réseau de neurones convolutif. Il rallonge surtout son temps d'entraînement. L'augmentation du nombre de classes d'objets complexifie le réseau dans les couches finales de neurones dites « fully connected » qui font le lien entre les dernières feature maps et les classes d'objets.

[708] Voir DenseCap: Fully Convolutional Localization Networks for Dense Captioning de Ku Fei-Fei, Stanford, 2015.

En 2015, des équipes de Google étaient capables de décrire une scène comportant plusieurs personnes et un objet (un frisbee), la prouesse étant d'ailleurs plutôt située dans l'agencement du réseau à mémoire **LSTM** que dans la reconnaissance des composantes de l'image[709]. Le système avait l'air de fonctionner dans une belle diversité de situations[710]. Depuis 2017, l'application de réalité augmentée mobile **Google Lens** ajoute des informations sur les objets visés par la caméra du smartphone, comme des fleurs et des restaurants[711].

L'état de l'art progressait grâce au défi **ImageNet Large Scale Visual Recognition Challenge** (ILSVRC) lancé en 2010 et renouvelé chaque année jusqu'en 2017[712]. Il permettait d'évaluer l'état de l'art de la reconnaissance d'images en mettant en concurrence plus d'une cinquantaine d'entreprises et laboratoires de recherche dans le monde. Le concours s'est arrêté une fois que les performances des réseaux de neurones étaient devenues plus que satisfaisantes pour des usages courants.

Les réseaux de neurones utilisés sont de plus en plus profonds (nombre de couche) et de plus en plus larges (nombre de catégories d'objets reconnus et taille des bases d'entraînement). Pour la reconnaissance d'images, on peut dépasser 150 couches de neurones, mais en général, toujours avec environ cinq couches de convolution.



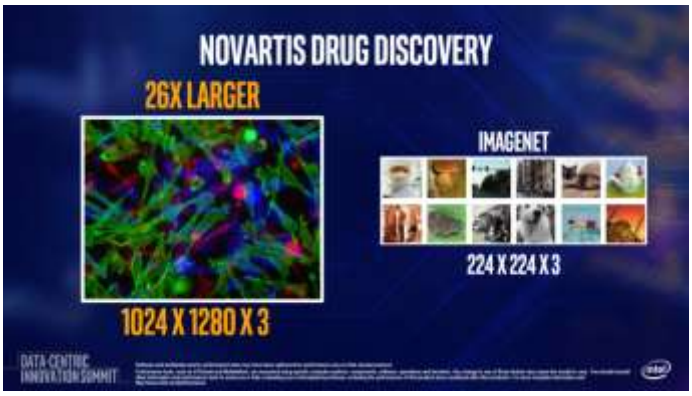
[709] Voir Show and Tell: A Neural Image Caption Generator, Google, 2015 (9 pages).

[710] Ce genre d'application peut exploiter la base Visual Genome qui est un jeu de données de 107 077 images dont les objets et régions sont labellisés avec des termes reliés dans des graphes. Voir Visual Genome Connecting Language and Vision Using Crowdsourced Dense Image Annotations, Stanford, 2016 (45 pages).

[711] Voir AI-powered Google Lens can identify types of flowers, give info about restaurants, mai 2017 et Voir Google Lens now recognizes over 1 billion products par Kyle Wiggers, 2018.

[712] Ce benchmark porte sur la reconnaissance d'images issue d'une base comprenant un million d'images dans 1000 classes différentes. Le niveau d'erreur mesuré est top-5 ou top-1. Le top-5 correspond à la proportion d'images pour lesquelles le bon label ne figure pas dans les cinq premiers considérés comme étant les plus probable par le réseau de neurones. Le top-1 correspond au label le plus probable. C'est le score le plus intéressant, le plus proche de la reconnaissance humaine. Le top-5 est un peu trop laxiste !

Des réseaux de neurones peuvent-ils analyser des images à haute résolution ? Cela semble être maintenant possible, en tout cas si l'on en croit Intel qui faisait de la reconnaissance d'image médicale sur une résolution de 1024x1280 pixels, en partenariat avec Novartis. Le tout avec des serveurs exploitant des chipsets Intel Xeon à 40 cœurs et surtout, une grande mémoire vive de 192 Go permettant de gérer les hyperparamètres du réseau de neurones.



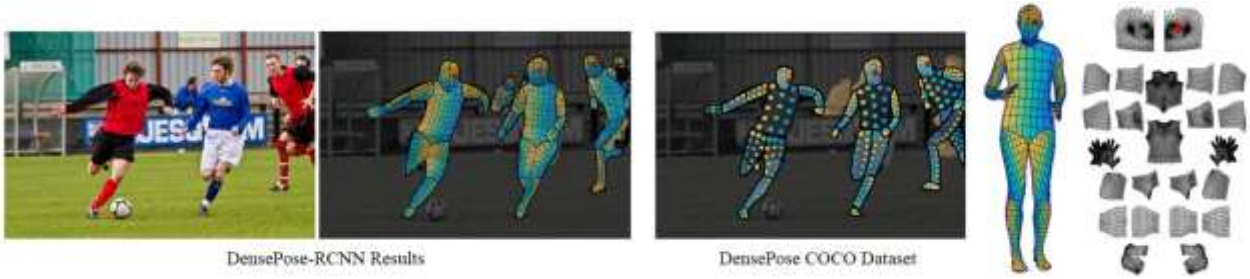
Le système pouvait ingérer 120 images de 3,9 mpixels par seconde pendant l'entraînement. Le tout tournait sur Tensorflow[713].

Lancé en open source début 2018, le projet **Detectron** de Facebook[714] associe un grand nombre de réseaux de neurones pour détecter et bien détourer les objets dans une image.

Il est développé en Python avec le framework Caffe2. Il utilise les réseaux de neurones Mask R-CNN, RetinaNet (tous deux créés en 2017), Faster R-CNN, RPN, Fast R-CNN, R-FCN et s'appuient sur ResNeXt, ResNet et un réseau convolutif VGG16. Cela illustre la notion d'intégration dans l'IA[715]. Pour obtenir une fonction donnée, il faut combiner jusqu'à une dizaine de techniques différentes !

Ces techniques sont exploitées dans les systèmes de conduite assistée et autonomes comme chez **Mobileye**, filiale d'Intel depuis 2017, qui détecte les piétons, les cyclistes, les autres véhicules, la signalisation au sol et les panneaux de signalisation avec de simples caméras RGB.

Début 2018, le projet **DensePose** réalisait une prouesse de plus en étant capable de bien isoler les différentes parties du corps dans la reconnaissance d'images de personnes et créant un modèle 3D du corps, allant au-delà des Mask R-CNN qui isolent au niveau des pixels les différents objets d'une image (*ci-dessous*)[716].



[713] Voir Using Deep Neural Network Acceleration for Image Analysis in Drug Discovery, mai 2018.

[714] Voir https://research.fb.com/downloads/detectron/. Cela fait suite à Fully Convolutional Networks for Semantic Segmentation 2016 (12 pages) issu de l'Université de Berkeley.

[715] Google propose une fonction de segmentation d'images équivalente, **DeepLab**. Elle est notamment employée dans les smartphones Pixel 2 et 2 XL pour flouter l'arrière-plan de portraits.

[716] Voir DensePose: Dense Human Pose Estimation In The Wild de Riza Alp Güler (Inria-Centrale-Supelec), Natalia Neverova et Iasonas Kokkinos (Facebook), 2018 (pages).

En avril 2018, Google faisait encore mieux en introduisant la reconnaissance de races de chiens dans l'application mobile **Google Lens** (source).

La reconnaissance d'images peut servir à la représentation des connaissances en s'appuyant sur une approche multimodale associant des descriptifs textuels riches d'images et les images proprement dites[717].

De nombreux types de réseaux de neurones permettent aussi de reconstituer la **profondeur des objets** dans une image sans même disposer de vision stéréoscopique[718].

En résumé, au-delà de la simple classification d'une seule image dans une catégorie, nous avons :

- La **segmentation sémantique** qui colorie les pixels d'une image en fonction des objets et zones détectés.
- La **classification et la localisation** qui détecte les objets et les entoure d'un cadre, avec un seul objet ou plusieurs objets.
- La **segmentation d'instances** d'objets différents[719].
- La **détection de la forme 3D** des objets à partir d'images en 2D[720] voire de plusieurs images d'un même objet pris sous plusieurs angles ou enfin, via une vue 3D obtenue par exemple avec un capteur de profondeur ou un LiDAR intégré dans une voiture[721].

Les progrès les plus récents concernent la détection des objets dans une vidéo. Elle opère maintenant en temps réel comme avec le framework open source **Yolo V3** que nous avons déjà cité[722]!

Malgré tout, il existe encore des zones d'ombre dans la reconnaissance d'images avec des cas particuliers présentant des effets de bord et des difficultés de reconnaissance pour les réseaux de neurones. Ce sont toutefois des cas assez marginaux[723].

L'interprétation des images est un pan entier de l'IA qui est la spécialité de nombreuses startups qui n'ont pas toutes été acquises par les GAFA ! Ces startups utilisent des techniques assez voisines basées sur le deep learning pour identifier le contenu de photos ou de vidéos, pour en extraire des labels qui sont ensuite exploitées dans diverses applications. En voici quelques-unes.

---

[717] Voir Multi-Modal Knowledge Representation Learning via Webly-Supervised Relationships Mining, 2018 (9 pages).

[718] Voir Deep learning for 3-D Scene Reconstruction and Modeling, 2017 (147 slides)

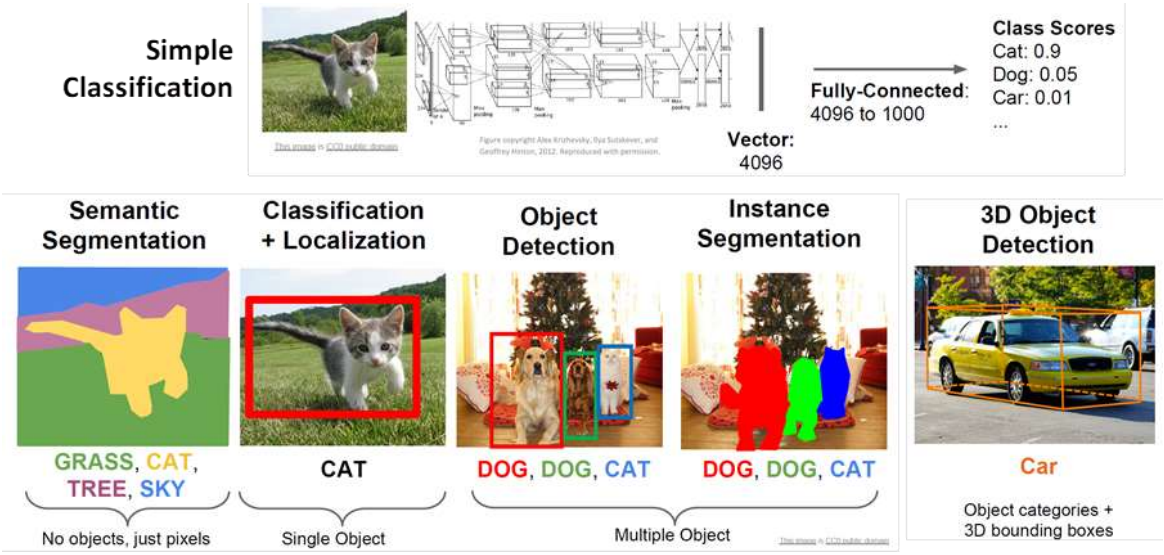
[719] Voir A Simple Guide to Semantic Segmentation par Bharath Raj, 2019.

[720] Voir FastDepth: Fast Monocular Depth Estimation on Embedded Systems, mai 2019.

[721] Voir le cours CS231N de Stanford de Fei-Fei Li, Justin Johnson et Serena Yeung et en particulier la onzième partie d'où est tirée l'illustration, dans Detection and Segmentation, mai 2018 (104 slides).

[722] Voir cet état de l'art : Deep Learning for Video Classification and Captioning, février 2018 (114 pages).

[723] Voir The world's most-advanced AI can't tell what's in these photos. Can you? Par Mark Wilson, juillet 2019.



Clarifai (2013, USA, $40M) propose une API en cloud permettant d'accéder à leurs fonctions de reconnaissance d'images. La startup a été créée par Matthew Zeiler, un ancien de l'équipe de Jeff Dean chez Google. Ils visent des marchés divers comme le e-commerce, les médias et le secteur public [724]. Leur solution intègre une fonction de recherche des les images.

Vicarious (2010, USA, $122M) est spécialisé dans la reconnaissance et la classification d'images. Ils se sont fait remarquer en étant capable d'interpréter des Captcha de toutes sortes avec une efficacité de 90%. Ils sont maintenant focalisés sur les usages de la reconnaissance d'image en robotique.

Cortica (2007, Israël, $69,4M) extrait les attributs clés d'images fixes ou animées pour les associer à des descriptifs textuels avec sa solution Image2Text.

Elle est par exemple capable de reconnaître une marque et modèle de voiture dans une vidéo ou un animal dans une photo. Cortica indique faire du deep learning non supervisé et temps réel. Le tout est protégé par 200 brevets ! La startup vise le marché de l'automobile, de la vidéosurveillance et des drones.

[724] Clarifai a ouvert en 2018 une filiale Neural Net One dédiée aux applications dans le secteur militaire et du renseignement. Ils travaillent notamment pour le Pentagone ainsi que pour les services de contrôle de l'immigration (ICE). Voir Despite a surge of tech activism, Clarifai plans to push further into government work, juin 2018.

Deepomatic (2014, France, $9M) utilise le deep learning pour interpréter le contenu, la forme et la couleur d'images dans les médias et les associer à des publicités contextuelles. Le tout pour de nombreux marchés (santé, transports, sécurité, industrie). Ils ont aussi développé une solution de checkout de restauration collective pour Compass, qui fait de la reconnaissance visuelle automatique du contenu des plateaux repas.

Le marché de l'inspection visuelle dans l'industrie comprend aussi **Neurala** (2006, USA, $20,1M) et sa solution embarquée Lifelong-DNN capable d'auto-apprentissage à la volée.

D'autres startups, notamment chinoises, se focalisent sur la détection d'activités en exploitant les images de caméras de surveillance. C'est ce que propose notamment **Umbo CV** (2014, Taïwan, $2,8M) qui s'appuie sur une solution logicielle fonctionnant dans le cloud.

Drone Volt (2011, France) commercialise via Aerialtronics un SDK pour sa caméra intelligente de drone Pensar. Cette dernière comprend un chipset Nvidia Jetson TX2 avec une capacité de calcul de 1,5 teraflops ainsi qu'un capteur HD et une caméra thermique infrarouge. Le SDK comprend 9 réseaux de neurones de reconnaissance d'images dédiés à la reconnaissance d'images aériennes.

Nalbi (2015, Corée du Sud) propose un outillage logiciel prêt à l'emploi pour la modification de photos avec extraction, copie, flou d'arrière et d'avant-plan. C'est en fait une bibliothèque logicielle de deep learning packagée servant au développement d'applications mobiles et embarquées.

Insightness (2014, Suisse) développe des systèmes embarqués de détection d'obstacles pour drones. Ils exploitent des capteurs photo Silicon Eye dotés de « SmartPixels » 320x262 pixels qui compriment l'information et détectent les mouvements, ce qui fait penser à ce que fait le Français Prophesee (vidéo). La solution est particulièrement économe en énergie. Le système peut aussi servir à alimenter des applications de réalité augmentée.

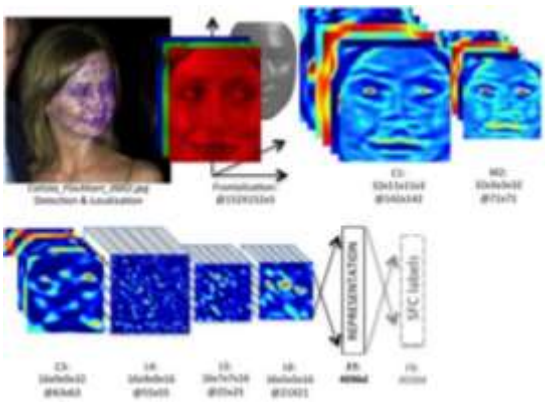
Des méthodes encore plus sophistiquées d'analyse d'images sont proposées qui exploitent les métadonnées de capteurs, comme celles de LiDARs ou arrivent à déterminer avec un réseau convolutif la focale de l'optique utilisée pour prendre une photo. Cela peut notamment servir à reconstituer des scènes 3D à partir de vues 2D[725].

Visages

La reconnaissance de visages est une technologie plutôt au point et depuis quelques années. Dans leur projet FaceNet, **Google** annonçait en 2015 avoir atteint un taux de réussite de détection de visage de 99,63%[726].

[725] Voir ce review paper : Convolutional Neural Networks With Heterogeneous Metadata par Patrick Langechuan Liu, mars 2020.

[726] Voir FaceNet: A Unified Embedding for Face Recognition and Clustering, juin 2015 (10 pages). Le tout s'appuyait sur un réseau neuronal à 22 couches avec comme entrée, des images de 220 pixels de côté.



De son côté, **Facebook** et son projet DeepFace s'appuyait sur la technologie issue d'une start-up israélienne **face.com** avec un taux de réussite de 97,25% pour vérifier qu'une personne sur une photo est la même sur une autre, quel que soit l'angle de la prise de vue et l'éclairage. C'est très voisin du taux de reconnaissance humain qui serait évalué à 97,5%.

On trouve de la détection de visages dans plein de solutions du marché comme avec la fonction Faces d'**Apple** iPhoto[727].

Les APIs en cloud des Cognitive Services de **Microsoft** apportent des services équivalents aux développeurs d'applications. Il en va de même pour **IBM Watson**, pour **Google** avec ses Cloud Vision APIs et **Amazon** avec son SDK Rekognition. Cette abondance des offres rappelle que les technologies de l'IA, une fois au point, deviennent rapidement des commodités. Les méthodes sont sur la place publique et facilement accessibles. Il faut ensuite les mettre en œuvre avec du logiciel et du matériel adapté aux besoins.

login par reconnaissance de visages
avec Windows Home et Intel RealSense

La différence se situe dans la mise en œuvre et aussi dans le marketing.

La reconnaissance des visages est évidemment un sujet chaud pour les services de sécurité. On en voit dans tous les films et séries TV ! En quelques secondes, les suspects sont identifiés. Est-ce comme cela dans la vraie vie ? Probablement pas. Cela explique pourquoi le **FBI** a lancé son projet NGI (Next Generation Identification) en 2009 qui est maintenant opérationnel. Il était pourvu à hauteur de la bagatelle de $1B et réalisé par **Lockheed Martin**.

Le marché de la reconnaissance faciale est aussi prolixe en solutions diffusées en OEM avec **imagga** (2008, Bulgarie, $642K) et ses API en cloud de tagging automatique d'images en fonction de leur contenu[728], **Cognitec (**2002, Allemagne) qui vise surtout les marchés de la sécurité et FaceMe de **Cyberlink**[729].

Nous avons aussi la startup française **Smart Me Up** (2012, France, 3M€, acquise par l'Italien Magneti Marelli en août 2018), qui propose une solution logicielle d'analyse des visages. Elle détecte l'âge, le comportement et les émotions des utilisateurs. La solution est commercialisée sous forme de brique logicielle en marque blanche utilisable dans des applications métier.

---

[727] Elle provient peut-être de la start-up suédoise **Polar Rose** acquise par Apple en 2010.

[728] Imagga Wordroom est un plugin d'Adobe Lightroom qui propose automatiquement des tags de labellisation des images à partir de la reconnaissance des objets, formes, couleurs et émotions. Les photos taggées sont analysées dans le cloud à partir de leur version en vignette. Le plugin est gratuit s'il est utilisé pour tagger moins de 2000 photos par mois.

[729] Voir FaceMe Moteur AI de reconnaissance faciale par CyberLink, 2020.

SenseTime (2014, Chine, $2,6B) commercialise une solution de reconnaissance de visages déclinée dans plusieurs verticaux dont le retail et les télécoms. Elle a été fondée par des chercheurs de Hong Kong.

Ses primitives fonctionnelles sont le suivi de plusieurs visages en temps réel dans des vidéos et la détection d'attributs divers (sourire, style de coiffure et barbe, âge, race, regroupements pour l'organisation d'albums photos, détection de visage vivant vs statique, maquillage virtuel). Elles visent le florissant marché de la vidéosurveillance en Chine, dans la distribution mais aussi l'imagerie médicale, la conduite autonome (avec Honda).

SenseTime est partenaire du MIT dans le programme de recherche Intelligence Quest et a Qualcomm ainsi qu'Alibaba parmi ses actionnaires. C'est l'une des startups de l'IA les mieux financées au monde. Ils disposent de leur propre centre de calcul avec 8000 GPU qui supporterait des réseaux de neurones avec des centaines de milliards de paramètres.

Megvii (2011, Chine, $1,4B) est un concurrent local de SenseTime avec son produit Face++ de détection de visage. Ces petits joueurs n'ont levé que $1,4B[730]. Hahaha ! Et ne parlons pas de **CloudWalk Technology** (2006, Chine, $389M) qui cible surtout la surveillance dans les aéroports.

En 2019, un chercheur chinois défrayait la chronique en présentant un prototype de caméra de surveillance de 500 mégapixels capable de reconnaître des dizaines de milliers de personnes d'un seul coup dans les lieux publics. La caméra est en fait un dispositif qui combine une cinquantaine de caméras de plus basse résolution. Elle doit exploiter un système d'assemblage des photos captées (*photo-stitching*), ce qui est facilité par le positionnement fixe des caméras les unes vis à vis des autres. Ensuite, un réseau de neurones doit analyser des portions d'images pour reconnaître les personnes[731]. Évidemment, la vie privée en prend un coup au passage, mais on est en Chine ! Tandis que la reconnaissance faciale est interdite à San Francisco ! Mais la vidéosurveillance, ce n'est rien par rapport à ce que présagent ces IA qui savent lire sur les lèvres[732].

On trouve des solutions de reconnaissance de visage dans les vidéos chez **Kairos AR** (2012, USA, $11M) qui savent aussi analyser les émotions et quantifier les foules[733], chez **KeyLemon** (2008, Suisse, $1,5M) qui propose une solution en cloud, chez **Matroid** (2016, USA, $33,5M), qui fonctionne sur des flux vidéo ou des photos ou chez le japonais **NEC**. Il faut aussi citer **OpenCV**, une solution open source de détection de visages et cela ne doit pas être la seule.

La reconnaissance de visages sert évidemment aussi aux applications de vidéo-surveillance, comme celle de **Camio** (2013, USA) qui fournit une solution en cloud d'exploitation de vidéos de caméras de surveillance.

---

[730] Megvii a recruté l'un des créateurs du réseau de neurones ResNet de Microsoft Research, Jian Sun, qui est directeur scientifique de la startup depuis 2016.

[731] Voir China's new 'super camera' can instantly pinpoint specific targets among tens of thousands of people par Tracey Shelton, septembre 2019. Cela provient de la Fudan University de Shanghai et du Changchun Institute of Optics, Fine Mecanics and Physics of Chinese Academy of Sciences de Changchun.

[732] Voir Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers par Ya Zhao, Rui Xu et al, novembre 2019 (9 pages).

[733] Voir Kairos gets a $4 million lifeline for its facial recognition software par Megan Rose Dickey, février 2019.

La reconnaissance de visage pourrait aussi servir à identifier les homosexuels avec une efficacité de 91%, ce qui pose d'évidents problèmes d'éthique dans les pays, régions ou systèmes politiques qui ne les tolèrent pas (genre Daech…) ou à même de générer diverses formes de discriminations comme aux USA[734]. Sans compter la difficulté de détection d'autres catégories (bi-sexuels, ...).

D-ID (2017, Israël, $22,9M) utilise des réseaux de neurones génératifs (GAN) pour modifier les visages stockés dans les bases de données clients. Cela rend ces images inutilisables par des hackers pour d'autres applications de reconnaissance faciale. Mais l'œil humain n'y verrait que du feu. Comment est-ce que cela fonctionne ? Ce n'est pas précisé. Le GAN doit au minimum ajouter un bruit imperceptible à la photo, si ce n'est ajouter quelques déformations légères[735].

Cela intéresse les entreprises qui stockent des photos de leurs collaborateurs ou clients. Cela serait un moyen de mieux respecter le RGPD en Europe.

C'est en tout cas une solution un peu plus élégante que ces maquillages conçus pour éviter la reconnaissance par les solutions de reconnaissance du visage, ou les méthodes voisines inventées par les manifestants à Hong Kong depuis l'été 2019[736].

DataSparQ (UK) propose AI Bar, un logiciel de reconnaissance de visage qui sert à optimiser l'attente dans les bars[737].

Une caméra surveille les gens qui s'approchent du bar et identifie les visages dans leur ordre d'arrivée. Ainsi, on peut servir les clients dans leur ordre réel d'arrivée.

Au passage, l'âge des clients serait vérifié. Plus de gruge possible ! Et en se grimant, on perdrait sa place dans la file d'attente (vidéo) ! En fait, la société à l'origine de cette application est un prestataire de service en data science qui a trouvé le moyen de faire parler d'elle avec cette solution originale.

Enfin, pandémie covid-19 oblige, l'IA est aussi mise à contribution depuis 2020 pour détecter les individus qui ne portent pas de masque par exemple pour un contrôle à l'entrée de lieux publics ou de bureaux[738].

---

[734] Voir 'I was shocked it was so easy': meet the professor who says facial recognition can tell if you're gay, par Paul Lewis dans The Guardian, juillet 2018. L'étude en question vient de chercheurs de Stanford : Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, par Yilun Wang et Michal Kosinski, 2017 (47 pages). La base d'entraînement utilisait 35 326 photos. L'efficacité de 91% est obtenue avec cinq photos par personnes. Mais tout cela reste à prendre avec des pincettes.

[735] Une autre technique consiste à ne pas enregistrer tous les points caractérisant le visage mais une partie seulement puis à utiliser des techniques de hashage de telle sorte qu'il devienne impossible de reconstituer le visage ou de l'identifier. Il reste possible de vérifier si le visage est bien reconnu.

[736] Voir Ce maquillage futuriste permet d'échapper aux systèmes de reconnaissance faciale par Nicolas Prouillac et Arthur Scheuer, juillet 2017.

[737] Voir Quand un outil de reconnaissance faciale permet de diminuer le temps d'attente dans les bars par (auteur pas trouvé), 2019.

[738] Voir Logiciel de détection de masque facial par Aerialtronics.

Émotions

L'analyse des visages, couplée ou non avec celle de la voix, permet de détecter des éléments extérieurs d'émotions humaines. Le principe de la reconnaissance des émotions dans le visage à partir de caméras est assez ancien. Elle est standardisée par le système de description **FACS** pour Facial Action Coding System[739], créé en 1978 par les psychologues américains Paul Ekman et Wallace Friesenen. Diverses startups se sont attaquées à cette fonction pour toucher divers marchés comme ceux du commerce et de la publicité mais aussi celui de la sécurité.

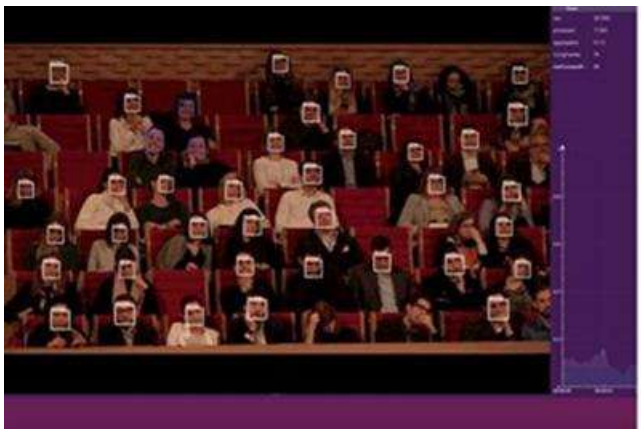
L'une d'entre elles est la startup **Affectiva** (2009, USA, $62,6M) que j'avais découverte au CES 2013[740]. Elle présentait une solution de captation des émotions d'un utilisateur exploitant une simple webcam sur un micro-ordinateur. Elle valorisait à l'origine un projet de recherche du MIT Media Lab et visait le marché de la publicité et du retail mais a eu du mal à le pénétrer. Ils s'intéressent depuis au marché de l'automobile pour détecter l'état du conducteur, comme le manque d'attention, qui est une fonctionnalité finalement assez limitée. Ils annonçaient avoir entraîné leurs modèles de deep learning avec 5 millions de visages issus de 75 pays[741].

Leur logiciel évalue les paramètres suivants : joie, gaieté, dégoût, mépris, peur, surprise, colère ainsi que la valence (état émotionnel allant du négatif au positif), l'engagement et l'attention. Le tout exploite l'analyse de 20 expressions faciales différentes via des réseaux convolutifs et des SVM. Mais il est difficile de savoir où cette solution est déployée d'un point de vue pratique.

En octobre 2019, **Fujitsu** annonçait avoir développé une solution équivalente avec l'Université de Carnegie Mellon et capable de déceler les micro-expressions plus efficacement et avec des images prises sous plusieurs angles de vue[742].

En France, la société **Datakalab** (2017) analyse simultanément plusieurs visages, comme les spectateurs d'un événement ou d'une conférence. Sa solution peut ainsi déterminer l'intérêt d'une audience pour une présentation, ou comparer cet intérêt entre deux intervenants comme en mai 2017 pendant le débat d'entre deux tours confrontant Emmanuel Macron à Marine Le Pen. Cela objectivisait une impression partagée sur la performance relative des deux finalistes ! L'opération était aussi réalisée sur une partie de l'audience de l'Echappée Volée début juillet 2018 à la Seine Musicale de Boulogne Billancourt pour comparer l'impact des différents intervenants.

Le service permet aussi d'évaluer le niveau de stress de clients, comme dans l'arrivée en gare. Datakalab se positionne comme un cabinet de conseil en neuromarketing. Il n'exploite pas que la vidéo mais aussi les informations issues de la voix et, optionnellement, de bracelets biométriques. Datakalab annonçait un partenariat avec IBM Watson en 2019.



---

[739] Ici : https://www.paulekman.com/wp-content/uploads/2013/07/Facial-Sign-Of-Emotional-Experience.pdf.

[740] Voir le Rapport CES 2013, page 256.

[741] Ils organisaient une conference intéressante en 2019, voir Emotion AI Summit 2019 Slides & Recorded Sessions par Affectiva, octobre 2019.

[742] Voir Fujitsu Develops AI based Facial Expression Recognition Technology to Accurately Detect Subtle Changes in Expression, octobre 2019.

Ils utilisent leur analyse d'émotion pour analyser le retour de clients lors de parcours dans des sites web afin d'optimiser ces derniers.

L'analyse des gestes et autres mouvements est un autre domaine où l'IA peut jouer un rôle. Il est pour l'instant moins courant que l'analyse des visages, mais se développe de plus en plus. Côté recherche, une étude européenne fait un état des lieux[743]. Elle illustre le fait que ce domaine est encore nouveau. L'équipe de recherche a mené une expérience avec un système à base de caméra et de Microsoft Kinect pour classifier des gestes et identifier les émotions associées. Elle met en avant le fait que la signification de ces émotions dépend de nombreux paramètres comme la culture des individus ainsi que le genre de la personne observée.

De son côté, **Google** a développé une API qui permet de détecter les gestes des mains avec précision, l'idée étant d'exploiter cela avec un smartphone, comme ils le font avec le Pixel 4 annoncé en octobre 2019[744].

Dans la même veine, **Sensing Feeling** (2016, UK, £436K) fait de la détection d'émotions à base de caméras pour analyser l'état d'une foule plus ou moins dense, utilisable par exemple dans le commerce (vidéo). Ca permettrait par exemple de savoir à quel étage du BHV à Paris les clients sont les plus heureux !

Avec, disent-ils, 95,76% de précision. Et c'est éthique par conception ! Pourquoi et comment ? Leurs capteurs n'enregistrent pas d'images et ne font pas de reconnaissance faciale.

Encore plus surprenants, ces travaux de recherche qui visent à détecter le niveau de stress d'utilisateur avec des caméras infrarouges analysant le rythme de respiration des humains avoisinants. Le tout exploite un simple réseau de neurones convolutif [745]!

Enfin, Apple faisait l'acquisition, début 2016, de la startup **Emotient**, spécialisée dans la reconnaissance d'émotions faciales à base de machine learning. Le matching de visages est une chose, mais détecter les émotions en est une autre et on peut s'attendre à ce qu'Apple utilise cette fonctionnalité dans les évolutions de ses solutions, notamment dans la visioconférence Facetime d'Apple.

La détection des émotions peut passer par d'autres intermédiaires que la captation vidéo des visages. Ainsi une équipe américano-iranienne a-t-elle développé un modèle de machine learning capable de détecter les émotions à partir d'électro-encéphalogrammes. Leur système associe trois méthodes de classification : les k-means, les SVM et des réseaux de neurones[746]. Bon, avec une base d'entraînement un peu maigrelette de 32 personnes qui avaient observé 40 vidéos musicales. Pas de quoi généraliser le processus à ce stade. Et heureusement, on n'a pas toujours un casque d'EEG sur le crane !

Tout ceci est évidemment à prendre avec des pincettes car la détection des émotions par analyse biométrique est partielle. Elle ne permet pas d'interpréter la nature exacte et l'origine des émotions. Il faut donc se garder d'utiliser ce genre de solutions à tort et à travers.

---

[743] Voir Survey on Emotional Body Gesture Recognition, publiée en janvier 2018 (19 pages).

[744] Voir Google open-sources gesture tracking AI for mobile devices par Kyle Wiggers, août 2019.

[745] Voir DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition using Low-Cost Thermal Imaging in Unconstrained Settings en août 2017.

[746] Voir Emotion Recognition with Machine Learning Using EEG Signals par Omid Bazgir & Al (équipe américaine et iranienne), 2018 (5 pages).

Imagerie médicale

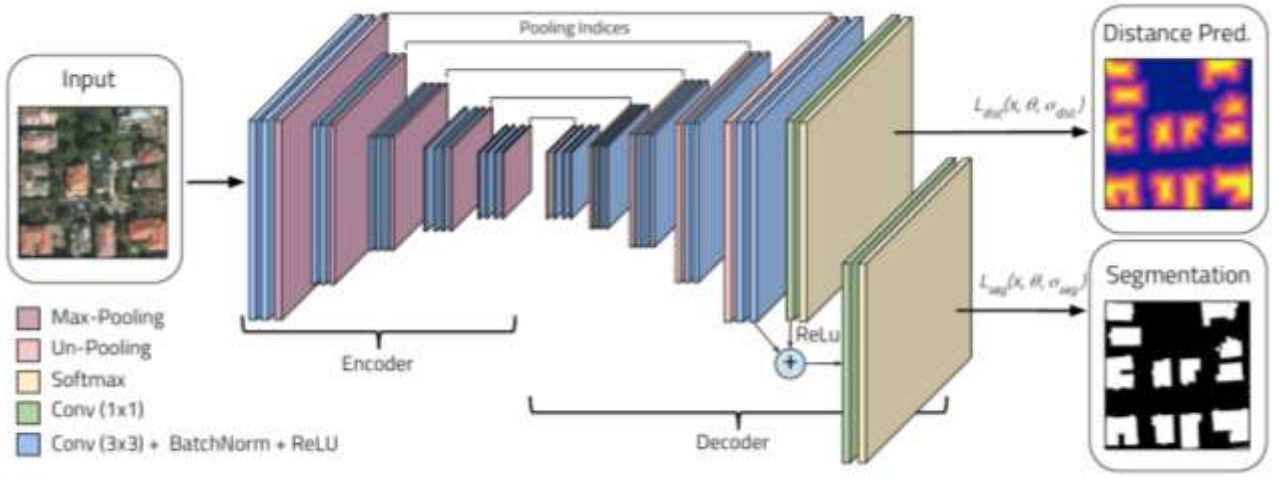
L'imagerie médicale est l'une des plus grosses applications du traitement d'images avec le deep learning. Elle est appliquée à la détection automatisée d'un grand nombre de pathologies au niveau de l'œil, de l'oreille, en dermatologie et en cancérologie. Elle exploite de nombreux réseaux de neurones différents en fonction des caractéristiques des images à traiter. Nous verrons cela plus en détail dans la rubrique sur la santé.

Télédétection

La reconnaissance d'images a aussi des applications en télédétection et imagerie satellite. Le deep learning permet de créer des solutions de recherche sémantique dans de gros volumes d'images, pour détecter des objets spécifiques comme des champs agricoles, des panneaux solaires ou des éoliennes, des plans d'eaux, des forêts, des toits, pour les caractériser en fonction de leur spectre lumineux, et pour analyser des variations dans le temps de ces paramètres. Cela sert aussi au fisc pour détecter des piscines !

Les innovations dans ce domaine comme ailleurs dans l'IA démarrent presque toujours au niveau de la recherche.

Ces différentes solutions utilisent une grande variété de réseaux de neurones. Ainsi, pour détecter la distance entre bâtiments et leur forme, on utilise un double *stacked auto-encoder* (*ci-dessous*).



Les applications touchent la détection d'objets en biais[747] et le décompte de véhicules[748], l'analyse de dommages sur des bâtiments, en particulier lors de conflits[749], l'analyse de dommages sur les toits après des intempéries[750], la détection de création de bâtiments dans des pays fermés comme l'Iran ou la Corée du Nord, le décompte de navires dans les ports et leurs mouvements et l'usage de données multi-spectrales couvrant le visible mais aussi l'infrarouge voir l'ultra-violet pour détecter le type de végétaux cultivés avec précision[751].

---

[747] Vue dans DOTA: A Large-scale Dataset for Object Detection in Aerial Images, 2018 (17 pages).

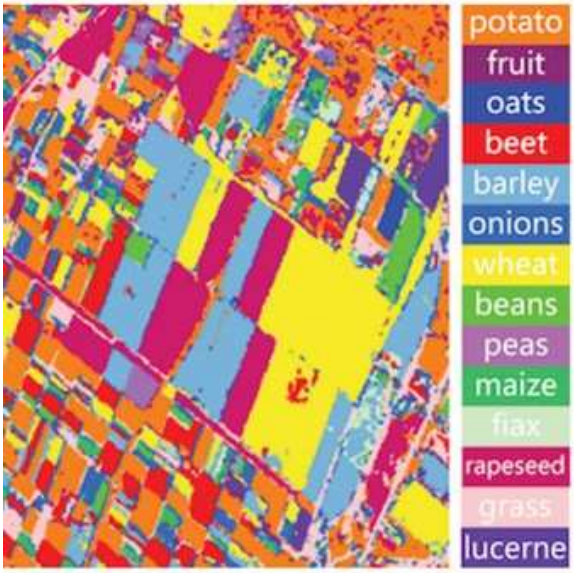
[748] Voir Real-Time Ground Vehicle Detection in Aerial Infrared Imagery Based on Convolutional Neural Network, 2018 (19 pages).

[749] Voir Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach, 2018 (8 pages).

[750] Voir Roof Damage Assessment using Deep Learning, 2018 (6 pages).

[751] Vu dans Deep Learning in Remote Sensing: A Review, 2018 (60 pages).

S'y ajoutent la préparation de la réponse face aux catastrophes naturelles, la détection d'inondations, l'analyse de déforestation [752], de l'impact d'incendies de forêts,[753] ou de la pollution atmosphérique.



Les applications commerciales sont aussi abondantes. En voici quelques-unes.

Airbus Defense and Space utilise le machine learning pour détecter les avions dans les aéroports avec de l'imagerie satellite. Ils les repèrent sur la base d'un jeu de données d'entraînement de 40 000 prises de vues avec la capacité à se débarrasser des nuages grâce à la mise en correspondance de plusieurs photos. Le taux d'erreur est inférieur à 4% (source).

Descartes Labs (2014, USA, $58M) exploite les données d'image satellite pour y découvrir comment évolue la production agricole, le cadastre des villes ou autres données géographiques, le tout via du machine learning développé sur TensorFlow et déployé sur Google Cloud.

Ils prédisent la production agricole à l'échelle mondiale ainsi que les risques de famine dans les pays émergents [754]!

TerraLoupe (2015, Allemagne, 3,2M€) analyse aussi les images satellite pour reconnaître ce qu'elles contiennent, en fonction des besoins clients (vidéo), comme analyser la surface des bâtiments dans le foncier, le type de toit, les antennes satellites, les panneaux solaires avec des applications dans l'agriculture, l'immobilier ou l'assurance (exemple *ci-dessus à droite*).

[752] Voir Emergency Response with Deep Neural Networks and Satellite Imagery, 2017 (76 slides) et Monitoring Deforestation in Rainforests Using Satellite Data A Pilot Study from Kalimantan, Indonesia, 2018 (26 pages).

[753] Voir cette thèse Making the most of machine learning and freely available datasets: A deforestation case study de Helen Mayfield, 2015 (368 pages).

[754] Voir cette vidéo montrant l'évolution dans le temps de la végétation aux USA et un exemple *ci-dessus à gauche* de détection automatique d'éoliennes, issu de cette vidéo.

C'est aussi l'activité de **Cape Analytics** (2014, USA, $31M) et d'**Orbital Insights** (2013, USA, $128,7M). Ce dernier décline son offre sur divers marchés, dont celui de l'énergie, pour évaluer les stocks et la consommation d'énergies fossiles en observant le niveau de profondeur des cuves de stockage[755] (*ci-contre*), ou celui du commerce pour identifier le trafic automobile dans les centres commerciaux et optimiser l'emplacement de nouveaux points de vente.

Comme Orbital Insights, **Earthcube** (2016, France, 4M€) propose aussi la détection d'objets via de l'imagerie satellite, applicable à des besoins de renseignement économique comme dans le marché des matières premières et des applications militaires pour la surveillance de sites stratégiques. Ils ont été repérés par le Ministère des Armées en France.

Contrôle qualité

Le contrôle qualité est très courant et se démocratise pour vérifier la qualité des pièces et produits fabriqués en usine. Il l'est aussi comme nous le verrons plus loin dans l'agro-alimentaire.

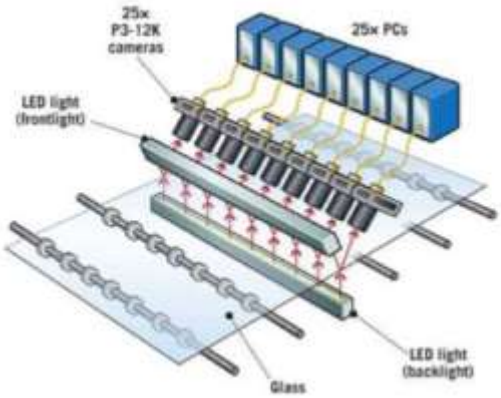
Le point clé de ces systèmes est qu'il doivent fonctionner en temps réel. Mais leur apprentissage est moins complexe car ils doivent analyser des images dont la variance est plutôt faible.

Le contrôle qualité utilisait le machine learning avant l'avènement des réseaux convolutifs !

Mais ces derniers ont permis de mettre en place plus facilement des solutions avec un entraînement automatique de réseaux de neurones.

Les solutions de contrôle qualité s'appuient sur l'exploitation d'imagerie dans le visible, l'infrarouge, les UV et même les rayons X. L'imagerie peut être complétée par d'autres types de capteurs comme ceux qui mesurent la variation de la résistance électrique de matériaux.

Tous ces systèmes doivent fonctionner en temps réel, au rythme de la fabrication dans les usines, d'où l'intérêt d'utiliser des caméras intégrant des processeurs neuromorphiques exploitant des réseaux de neurones déjà entraînés. Voir d'autres solutions du genre dans la partie fabrication et contrôle qualité du marché « industrie ».



[755] Voir https://orbitalinsight.com/products/energy/#slider-5.

Hacarus (2014, Japon, $1,2M) cible l'industrie et la santé avec une solution d'inspection visuelle qui détecte les caractéristiques des objets avec de très faibles volumes de données d'entraînement (moins d'une centaine d'images pour détecter des défauts) et le fait de manière « explicable ». Et c'est loin d'être la seule startup à faire ce genre de promesse.

Qualité d'images

Des réseaux de neurones peuvent aussi servir à détecter la qualité d'images. C'est ce que fait **Imginit** (2017, France), en analysant la qualité d'images, pour l'acquisition d'images tierces parties, notamment dans les médias, la publicité, la gestion de stocks de photos.

Regaind (2014, France, 400K€) est une startup qui proposait une solution de tri automatique de photos en cloud s'appuyant sur diverses méthodes de machine learning et de deep learning.

Elle permettait de trier les photos sous un angle à la fois narratif et descriptif et de les tagger automatiquement. Elle comparait diverses caractéristiques des photos : leur cadrage, le flou d'arrière-plan, les couleurs, etc. La startup a été acquise par Apple pendant l'été 2017 et depuis, c'est le trou noir. La fonction pourrait apparaître dans iPhoto, si ce n'est pas déjà fait.

Google Photos propose aussi une fonction équivalente de tri de photos.

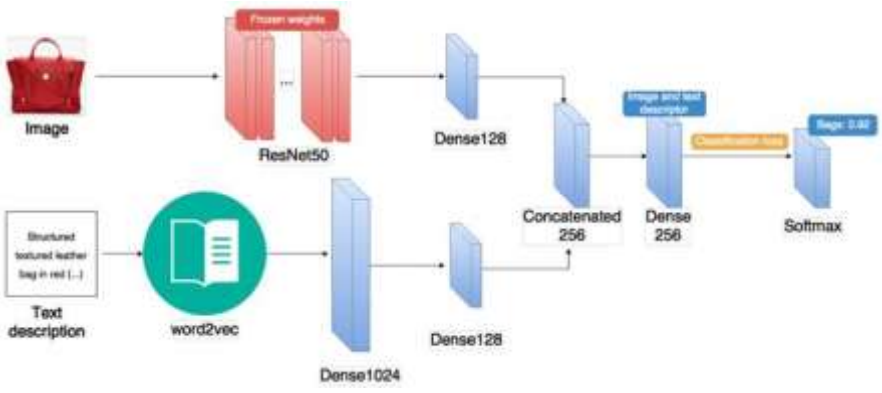
En décembre 2018, **Xiaomi** présentait DeepExposure, un prototype de solution exploitant des GANs permettant de corriger les problèmes d'exposition et de contraste de photos prises avec des smartphones et améliorant au passage les détails[756].

Vous trouverez d'autres exemples d'usages de l'IA dans le secteur de la photo dans la rubrique associée dans le marché des médias et des contenus.

Recherche

Les moteurs de recherche sont de gros utilisateurs de machine learning, en particulier pour l'indexation et la recherche d'images. La création de métadonnées d'images provient à la fois des données des pages web qui les contiennent et de leurs noms de fichiers, mais elles peuvent être enrichies par la reconnaissance d'objets à base de deep learning.

Un moteur de recherche d'objets peut retrouver des images à partir de leur description textuelle en associant plusieurs types de réseaux de neurones (*ci-contre*)[757]. En juillet 2018, **Microsoft** lançait une fonction de recherche similaire à celle de Google Lens dans Bing.



[756] Voir Xiaomi details DeepExposure, an AI that automatically fixes image exposure and detail, par Brittany Hillen, dans DpReview, décembre 2018. Et le papier d'origine : DeepExposure: Learning to Expose Photos with Asynchronously Reinforced Adversarial Learning, 2018 (11 pages).

[757] Voir DeepStyle: Multimodal Search Engine for Fashion and Interior Design, janvier 2018 (11 pages).

Elle permet de trouver un objet dans les sites de vente en ligne à partir d'une photo captée par son smartphone[758].

Quelques startups proposent des solutions de recherche d'images pour les entreprises. Mais ces fonctions sont largement utilisées par les grands acteurs du Web et du mobile manipulant des images, tels **Facebook**, **Instagram**, **Pinterest**[759], **eBay**[760] et **Amazon**.

La recherche d'images similaires est aussi courante. Elle sert à la recommandation dans les systèmes de vente en ligne, dans les moteurs de recherche ainsi qu'à l'identification de contrefaçons de produits de marques. Elle s'appuie aussi sur du deep learning. Dans ce créneau, l'application mobile **Gradient** permet de retrouver les stars à qui l'on ressemble et de retoucher sa photo, aux prix d'un abonnement mensuel à la limite du scam[761].

Caractères et écriture

La reconnaissance de caractères est réalisée dans les textes issus de scans (OCR) ou issus de prises de vue en temps réel (caméras, véhicules autonomes). Les systèmes actuels savent détecter les textes, les images et schémas de documents scannés. Nous avions même un leader en France dans le domaine avec la société **LTU**, acquise par le japonais **Jastec** en 2005 puis redevenue indépendante en 2019.

La reconnaissance de l'écriture manuscrite est aussi réalisée à la volée, à partir d'encre digitale, saisie par exemple avec un stylet comme sur les tablettes. Ce marché est moins connu que pour la reconnaissance vocale ou d'images. Et nous y avons un champion français avec la société **MyScript** (1998, France), anciennement Vision Objects, qui est basée à Nantes et qui a notamment commercialisé son logiciel auprès de **Samsung** pour ses Galaxy Note.

Nous avons aussi le californien **Vilado** (2011, USA, $52M), qui extrait les informations de l'écriture manuscrite et convertit automatiquement les formulaires en tableaux avec des applications évidentes dans les assurances et toutes les bureaucraties imaginables (vidéo). La solution en cloud est même reliée à Salesforce.

---

[758] Voir Microsoft launches AI-powered Bing Visual Search, juillet 2018 (vidéo).

[759] Voir Visual Search at Pinterest, 2015 (10 pages).

[760] Voir Visual Search at eBay, 2017 (10 pages) qui évoque l'utilisation d'un Convnet ResNet pour la classification d'images.

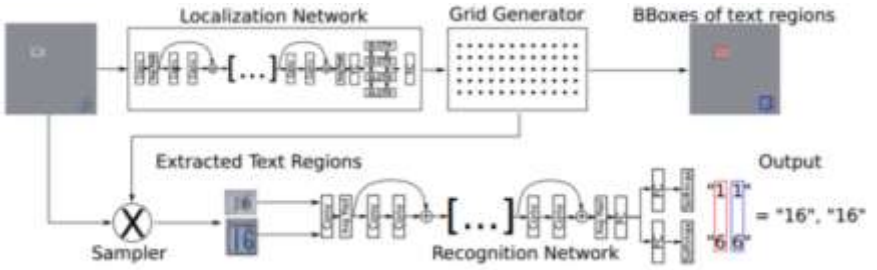
[761] Voir Restez sur vos gardes en utilisant Gradient, l'appli qui révèle à quelle star vous ressemblez par Albane Guichard, octobre 2019.

La reconnaissance d'écriture est d'ailleurs réversible car on peut aussi générer de l'écriture manuscrite synthétique à partir d'une écriture existante comme dans ce projet de recherche de l'Université de Toronto (*ci-contre*). C'est moins connu que les voix synthétiques[762].

Même si elles font moins les choux gras de l'actualité sensationnaliste de l'IA, les techniques d'OCR continuent de progresser. Les variantes portent sur les réseaux de neurones utilisés : CNN (réseaux convolutifs pour détecter les blocs de textes dans les pages, puis les mots et les caractères), LSTM (réseaux à mémoire, pour reconstituer des mots, expressions ou phrases qui se tiennent) et autres RNN (réseaux récurrents)[763].

Les solutions de reconnaissance d'écriture portent sur de l'écriture imprimée, le cas le plus simple, jusqu'à l'écriture manuscrite.

La reconnaissance d'écriture imprimée porte aussi sur les menus dans les restaurants aussi bien que sur les panneaux de circulation pour les véhicules autonomes (*ci-contre*[764]).



D'autres solutions se spécialisent dans l'analyse globale de documents manuscrits historiques[765].

Réseaux de neurones génératifs

Les réseaux de neurones générateurs de contenus sont apparus en 2014 grâce aux travaux de Ian Goodfellow de l'Université de Montréal avec une équipe comprenant aussi Yoshua Bengio[766].

---

[762] Voir aussi Synthetic handwritten text generation par Adria Rico Blanes, 2018 (14pages).

[763] Voir Build a Handwritten Text Recognition System using TensorFlow par Harald Scheidl, 2019, Handwritten Text Recognition using Deep Learning, par Batuhan Balci et al, 2016 (8 pages), Optical Character Recognition using Deep learning – A Technical Review par Preeti Bhatt, juin 2018, qui porte sur l'amélioration de la reconnaissance des caractères. Et puis Using ASR methods for OCR par Ashish Arora et al, 2019 (6 pages) qui s'inspire de la reconnaissance de la parole et de la traduction automatique pour associer différents types de réseaux de neurones permettant de repérer les mots hors-vocabulaires. Ils associent des TDNN (time delay neural network) des CNN (réseaux convolutifs). Deep Learning for Word Spotting par Gernot Fink, 2018 (106 slides) décrit différentes méthodes de détection de mots manuscrits dans un texte et l'assemblage de réseaux de neurones de types différents pour y parvenir. A Scalable Handwritten Text Recognition System par R. Reeve Ingle et al, juin 2019 (8 pages) améliore l'efficacité de la reconnaissance d'écriture avec de gros modèles de données d'entraînement et avec des réseaux de neurones de type LSTM. Handwritten Text Segmentation via End-to-End Learning of Convolutional Neural Networks par Junho Jo, juin 2019 (6 pages) s'attaque à l'écriture manuscrite avec des CNN au lieu de LSTM ainsi qu'à des documents mixtes contenant du texte imprimé annoté par écriture manuscrite. Improving CNN-RNN Hybrid Networks for Handwriting Recognition par Kartik Dutta et al, 2018 (6 pages) s'appuie sur un mix de CNN et de RNN pour la reconnaissance d'écriture manuscrite.

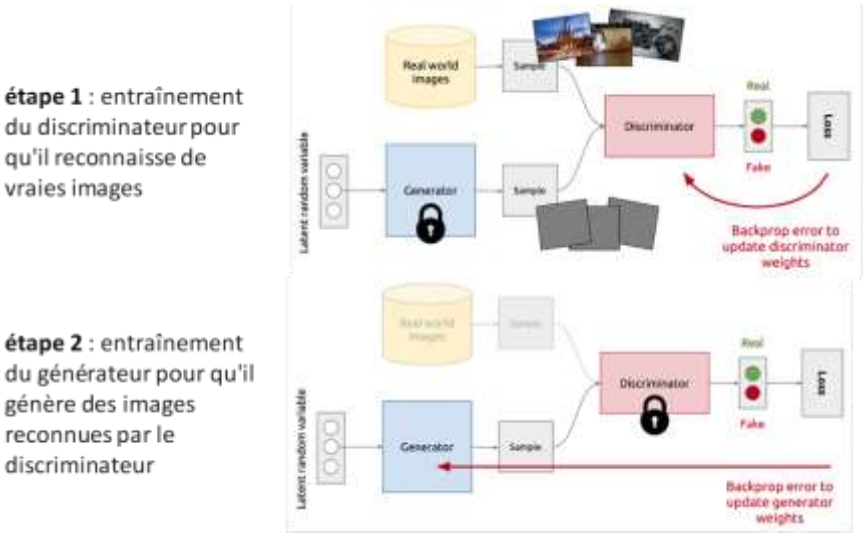
[764] Voir STN-OCR: A single Neural Network for Text Detection and Text Recognition, 2017, exploite un premier réseau de neurons de détection de régions de textes dans une image, puis un autre qui fait la reconnaissance elle-même. Il sert surtout à reconnaître les panneaux de noms de rues.

[765] Voir Start, Follow, Read: End-to-End Full-Page Handwriting Recognition par Curtis Wigington & Al, 2018 (17 pages) qui s'attaque à l'OCR de documents historiques anciens complets.

[766] Voir Generative Adversarial Nets, 2014 (9 pages).

Ces « GAN » impressionnent par leurs capacités à « prédire » l'univers visuel à partir de peu d'informations. Ils complètent des images ou les transforment et génèrent des images assez plausibles pour le cerveau humain. Au point que l'on en vient à trouver que l'IA est créative. Mais elle ne fait dans ces cas qu'appliquer des algorithmes de la même manière et sans discernement, comme n'importe quel filtre de retouche d'images dans Photoshop. Et elle exploite des éléments de créativité d'origine humaine. Faut-il donc lui coller des attributs de créativité comme celle des artistes ? Pas encore, même si certains croient que ce temps est déjà venu[767].

Les GANs peuvent servir à générer d'autres types de contenus et notamment du texte, ce que nous verrons dans la partie associée. D'un point de vue pratique, les GANs associent un générateur de contenu et un discriminateur. Le générateur crée un contenu artificiel extrapolé à partir de contenus existants et le discriminateur vérifie que le contenu est crédible vis-à-vis d'un jeu d'entraînement.



Le générateur est un réseau de neurones convolutifs de type stacked autoencoder qui génère un contenu à partir d'un autre contenu, éventuellement moins dense en information[768]. On ajuste son fonctionnement avec des paramètres divers selon les cas. Le générateur et le discriminateur se concurrencent l'un et l'autre pour ajuster le contenu généré pour qu'il soit bien reconnu comme « crédible » par le discriminateur.

Il existe en fait une très grande variété de GANs. Avinash Hindupur en a répertorié plus de 420[769]. On y trouve par exemple les VAE (Variational Autoencoder), les DCGAN et WGAN (Wasserstein Generative Adversarial Network). Comme tous les réseaux de neurones, les GANs sont des outils probabilistes. Ils ne fournissent pas « la réponse » ou « une réponse parfaite », mais une réponse plausible parmi d'autres, et dont la plausibilité est sujette à caution.

L'entraînement des GANs est difficile. On doit disposer de bons jeux de données pour les alimenter[770]. La variété des GANs s'explique par le fait qu'ils réalisent des fonctions très spécialisées, au coup par coup. Il n'existe pas de GAN généraliste capable de tout faire. C'est un problème homothétique avec celui du Graal de l'AGI, l'intelligence artificielle généralisée.

---

[767] Voir The Coming Creativity Explosion Belongs to the Machines, de Melba Kurman, octobre 2017 qui confond comme c'est courant la créativité des machines et celle des hommes qui les ont programmées.

[768] Voir Generative Adversarial Networks de Ole-Johan Skrede de l'Université d'Oslo, mai 2018 (88 slides) ainsi que Generative Adversarial Networks (GANs) de Binglin, Shashank et Bhargav, 2017 (73 slides) qui expliquent bien le procédé. Cette dernière source est celle des schémas de cette page.

[769] Dans The GAN Zoo, qui était actualisé en février 2018.

[770] Voir Limitations of Encoder-Decoder GAN architectures, de Sanjeev Arora et Andrej Risteski, mars 2018, qui décrit bien les limitations des GANs et leur papier Do GANs learn the distribution ? Some theory and empirics, 2018 (16 pages). Vous avez aussi quelques bons exemples de ratés dans This AI is bad at drawing but will try anyways, août 2018.

Les exemples de GANs abondent et nous allons faire ici le tour de quelques-uns des plus marquants d'entre eux[771]. Nous verrons au passage que les travaux de recherche dans les GANs sont vite récupérés par les startups ou les grands du numérique comme Google.

Les GANs ont un autre usage intéressant : ils peuvent aussi servir à générer des jeux de données pour entraîner des réseaux de neurones de reconnaissance d'images, notamment dans l'imagerie médicale[772] !

Colorisation

Il existe des GANs qui savent automatiquement coloriser des images en noir et blanc. Cela peut servir à moderniser des contenus anciens, souvent argentiques, aussi bien photo que cinéma.

Il faut évidemment de grandes bases d'entraînement pour que les couleurs soient les bonnes. Les exemples présentés dans les publications scientifiques sont toujours plausibles mais il doit sûrement y en avoir qui ne fonctionnent pas bien lorsqu'il peut y avoir ambiguïté sur une couleur. Un ciel bleu est en fait peut-être gris, même si la couleur du ciel peut avoir un impact sur le reste d'une photo de paysage.

Colorful Image Colorization, Zhang, Isola & Efros, 2016

Google Photos, mai 2018

Plusieurs GANs ont été publiés pour la colorisation depuis 2016[773]. En mai 2018, **Google** annonçait l'intégration de cette fonctionnalité dans son logiciel mobile Google Photos lors de la conférence Google I/O, parmis d'autres qui font aussi appel au deep learning (vidéo) comme la colorisation sélective d'images !

En 2019, **Adobe** annonçait intégrer cette fonction de colorisation dans Photoshop Elements 2020 et Premiere Elements 2020[774].

Dans le même registre, les GANs peuvent aussi servir à transformer des photos en bande dessinée, sans les légendes humoristiques pour l'instant. En voici un exemple récent réalisé en Chine (*ci-dessous*, avec les architectures en couche de son discriminateur et de son générateur)[775].

---

[771] Certains d'entre eux sont issus de cette présentation : Generative Adversarial Network (GAN), de Hongsheng Li (non daté, 88 slides).

[772] Voir par exemple Training artificial intelligence with artificial X-rays, juillet 2018.

[773] Comme Colorful Image Colorization de Richard Zhang, Phillip Isola et Alexei A. Efros de Berkeley 2016 (29 pages).

[774] Voir Adobe Photoshop and Premiere Elements 2020 arrive with new AI-powered tools par Brittany Hillen, octobre 2019.

[775] Vu dans CartoonGAN: Generative Adversarial Networks for Photo Cartoonization, 2018 (10 pages),

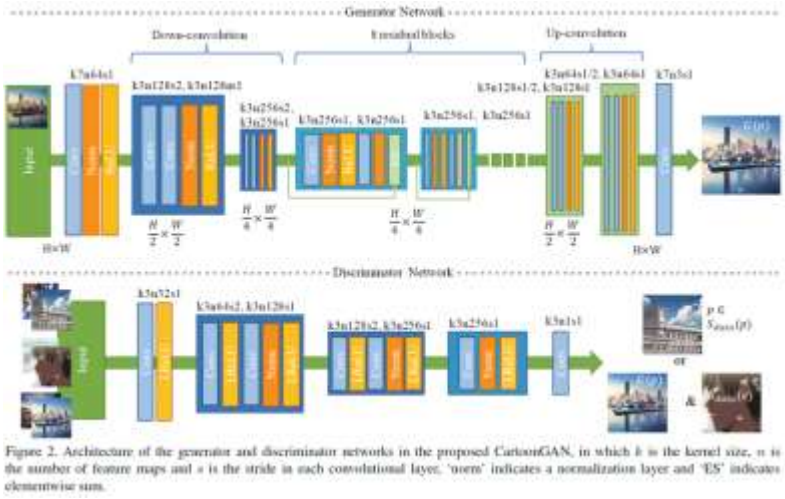


Figure 2. Architecture of the generator and discriminator networks in the proposed CartoonGAN, in which k is the kernel size, n is the number of feature maps and s is the stride in each convolutional layer. 'norm' indicates a normalization layer and 'ES' indicates elementwise sum.

Amélioration

Les GANs et/ou les stacked autoencoders permettent aussi d'améliorer des images, notamment celles qui sont bruitées. C'est ce qu'a notamment démontré Nvidia en juillet 2018 en collaboration avec des chercheurs de l'Université de Aalto en Finlande et du MIT aux USA[776].

Leur solution utilise des GPU Nvidia Tesla P100 et un logiciel développé avec le framework de réseau de neurones cuDNN exploitant le jeu d'instructions CUDA des GPU Nvidia. Voir la vidéo associée qui contient diverses démonstrations avec des reconstitutions réalistes de portions supprimées de l'image.

Figure 2. Random impulse noise. Our denoiser is trained on corrupted image pairs only.

Le système a été entraîné avec 50 000 images de la base ImageNet. Cela pourrait servir dans l'imagerie médicale et en astronomie. Mais comme l'indique un commentaire dans un article de Dpreview[777], les jeux de tests utilisés étaient à basse résolution et consommaient énormément de ressources machine. La généralisation d'un tel procédé sur des photos de reflex à haute résolution pourrait donc attendre quelques cycles de la loi de Moore !

D'autres usages des GANs ont des applications grand public plus directes comme l'adaptation de l'exposition ou même de l'éclairage des photos.

Ce genre de fonction est cependant de plus en plus souvent intégré dans les fonctions photos des smartphones. Elles exploitent les NPU (Neural Processing Units) qui sont intégrés dans les chipsets des smartphones.

Une telle fonctionnalité a été développée par des chercheurs d'Inria à Sophia Antipolis avec Adobe et l'UC Berkeley dans le projet **DeepRelighting**[778]. Elle permet de modifier les conditions d'éclairage de photos et de vidéos via un réseau de neurones entraîné avec un jeu d'images de synthèse.

---

[776] Voir Noise2Noise: Learning Image Restoration without Clean Data, 2018 (12 pages).

[777] Sur NVIDIA researchers develop AI that removes noise from images with incredible accuracy, juillet 2018.

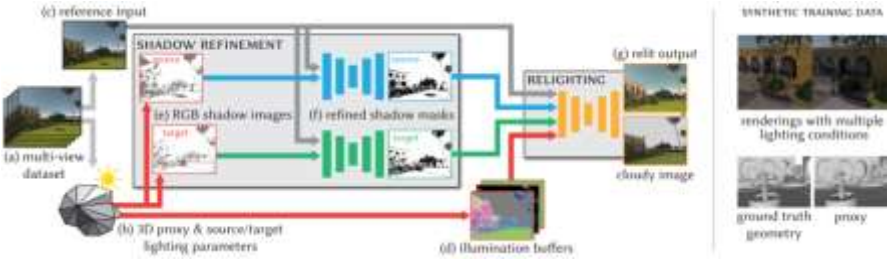
[778] Voir Multi-view Relighting Using a Geometry-Aware Network, 2019 (vidéo).

On peut prendre une photo à midi et la transformer en photo prise en fin de journée. Le système exploite un système de reconstruction d'un modèle 3D de la scène pour ensuite l'éclairer de manière différenciée et même gérer les ombres.

Figure 1: Une photo de Manarola prise en milieu de journée.

Figure 2: La même image éditée grâce à la méthode développée chez Inria.

Le système a été entraîné avec un grand nombre de jeux de données synthétiques issues de modèles 3D. Cette fonctionnalité est intégrée dans Photoshop Camera, annoncé en novembre 2019[779].



Enfin, **Swapping Autoencoder** est un GAN qui permet de manipuler des images au niveau des textures, comme dans les exemples ci-dessous, pour modifier les briques d'une cathédrale ou le ciel[780].

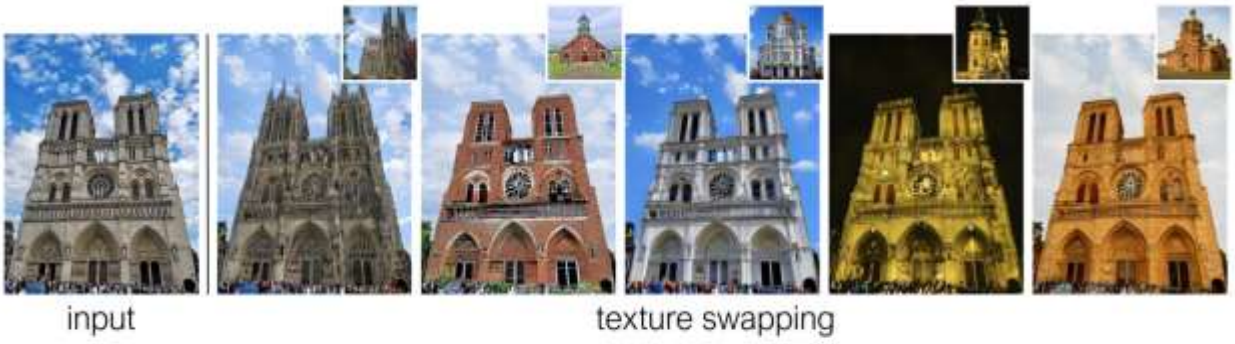


Figure 1: Our Swapping Autoencoder learns to disentangle texture from structure for image editing tasks. One such task is texture swapping, shown here. Please see our project webpage for a demo video of our editing method.

Résolution

L'amélioration de la résolution d'images génère aussi des résultats étonnants[781]. L'idée aurait été proposée la première fois par Ian Goodfellow en 2016.

[779] Voir la vidéo Adobe MAX 2019 Opening Keynote - Accelerating Your Creativity, novembre 2019, à partir de 2h18 mn.

[780] Voir Swapping Autoencoder for Deep Image Manipulation par Taesung Park et al, décembre 2020 (23 pages).

[781] Voir Super resolution with Generative Adversarial Networks de Boris Kovalenk.

Elle est depuis déclinée dans de nombreuses productions de chercheurs qui veulent rendre leurs méthodes les plus génériques possibles[782]. Les variantes sont très nombreuses pour améliorer la texture des images[783]. Cela rappelle par la même occasion le scénario de l'excellent film « Sens Unique » avec Kevin Costner, réalisé pendant les années 1980 et qui voit la NSA faire ce genre de reconstitution.

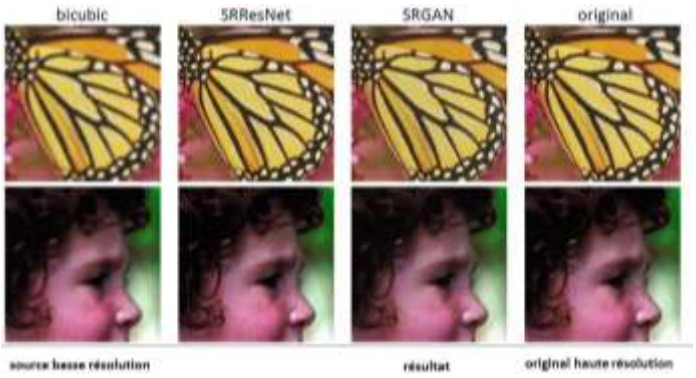


Figure 1. Low-resolution blurry images are challenging for the state-of-the-art super-resolution and deblurring methods. Sequentially applying super-resolution and deblurring methods further exacerbates the artifacts. Our method learns to reconstruct realistic results with clear structures and fine details.

La startup **Let's Enhance** (2017, Estonie) propose d'upscaler des photos à la demande[784]. Cette fonctionnalité est aussi proposée par **Skylum Software** (2008, USA) dans son logiciel Luminar AI qui améliore les textures de vos photos[785]. Une fonction d'amélioration de la résolution de photos est aussi intégrée dans **Adobe Lightroom** depuis début 2019. Elle est intégrée dans le dématriçage de Bayer dans le traitement des fichiers RAW et a été entraînée avec plus d'un milliard d'images de référence.

La méthode est aussi mise en œuvre dans l'upscaling de vidéos de basse résolution en plus haute résolution, par exemple pour passer du 1080p à la 4K ou à de la 8K, ou pour restaurer des films anciens. Les méthodes et jeux de données d'apprentissage utilisés par ce genre de GAN s'appuient sur des vidéos 4K récupérées sur YouTube et réduites en 720p (1280x729 pixels). Ils en extraient 300 000 carrés de 128 pixels de côté dont ils créent des versions réduites de 32x32 pixels qui servent à entraîner le système. Je simplifie un peu mais cela reste impressionnant, même si la présentation doit ne conserver que les cas où cela fonctionne bien, appliquant le biais du survivant à la démonstration[786]! On ne voit que ce qui fonctionne et pas ce qui ne marche pas, biaisant notre appréciation du potentiel d'utilisation en production de ces algorithmes.

---

[782] Voir Photo-realistic single image super-resolution using a generative adversarial network, de Christian Ledig et son équipe, 2016 (19 pages) qui est cité dans NIPS 2016 Tutorial: Generative Adversarial Networks, 2016 (57 pages).

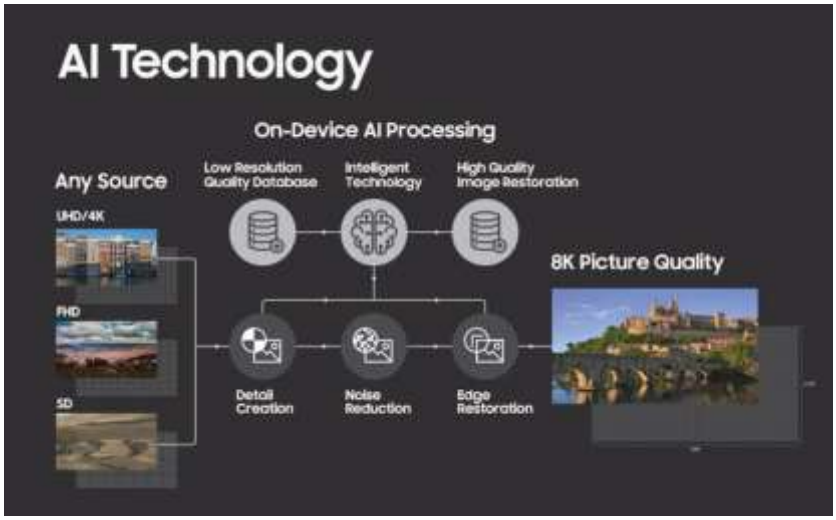
[783] Voir A Fully Progressive Approach to Single-Image Super-Resolution (10 pages), EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis, 2017 (19 pages) qui améliore la texture des images et Learning to Super-Resolve Blurry Face and Text Images (10 pages) qui part d'une photo vraiment très floue pour générer un visage proche de la vérité (*ci-dessous*). Magique ? Non, ce sont juste des modèles probabilistes. Les méthodes d'upscaling de photos sont régulièrement améliorées. Voir par exemple PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models par Sachit Menon et al, juillet 2020 (vidéo). Ce système avait cependant généré une polémique de plus sur le biais des algorithmes. En effet, un visage de personne issue de minorités visibles (aux USA) avait été upscalé vers un personnage blanc. Voir cette discussion sur Twitter. Encore une fois, c'était probablement lié à une base d'entraînement ne comportant pas suffisamment d'images.

[784] Elle propose un test d'upscaling, permettant l'upload d'une photo, mais demande ensuite les coordonnées de l'utilisateur. Je ne l'ai donc pas testée. La génération de leads à la petite semaine me fatigue !

[785] Voir Luminar's AI Structure selectively enhances textures and detail in your images par Lars Rehm, août 2019. Luminar AI propose d'autres fonctionnalités à base d'IA qui gèrent plus ou moins automatiquement les paramètres d'ajustement de photos.

[786] Voir Photorealistic Video Super Resolution, juillet 2018 (15 pages).

Le plus étonnant était de retrouver cette technique au CES 2018 chez Samsung qui introduisait cette fonction dans un prototype de TV 8K, capable d'upscaler des contenus SD, HD et UHD en 8K. La méthode utilisée doit être voisine, utilisant un réseau de neurones entraîné à upscaler des morceaux d'images. Mais on attend de voir ce que cela donne à grande échelle. Cela doit bien entendu dépendre de la nature des bases d'entraînement[787].



L'upscaling de vidéos et films a par exemple permis de restaurer des films des missions Apollo ainsi que la légendaire arrivée du train à la Ciotat de 1896[788].

Une solution de ce type fonctionnant en mode batch pour les producteurs de vidéos est proposée par **Topaz Labs** (2005, USA) pour $200, en plus d'une suite équivalente pour le traitement de photos, dénommée Sharpen AI.

En 2020, **Nvidia** démontrait comment utiliser l'upscaling pour améliorer la qualité des visio-conférences qui sont souvent transmises dans une résolution dégradée par les Zoom et autres Microsoft Teams et Google Meet[789].

Compléments

Le quatrième grand exercice de style des GANs est de compléter des images incomplètes. C'est une sorte d'exercice de prévision.

Les résultats sont intéressants mais imparfaits, surtout si l'on regarde de très près les images générées, qui par ailleurs sont à basse résolution.

Il ne faut pas oublier que ce sont des modèles probabilistes ! Et lorsque l'effet est de petite taille, comme à droite, vous n'y voyez que du feu[790] ! Cela joue sur les limites de notre cortex visuel qui lui aussi interprète les images générées par la rétine par approximation.

[787] La méthode de superrésolution est décrite en détails dans An Introduction to Super Resolution Using Deep Learning, par Bharath Raj en août 2019. Voir aussi Image Analysis and Synthesis Deep Learning use cases at Technicolor, 2018 (40 slides).

[788] Voir AI upscales Apollo lunar footage to 60 FPS par Nancy Atkinson, juillet 2020, qui upscale des vidéo et les passe de 12 à 60 images par seconde (vidéo) et How AI helped upscale an antique 1896 film to 4K par Andrew Tarantola, février 2020 (vidéo).

[789] Voir NVIDIA Turns To AI To Improve Video Conference Calls par Tyler Lee, mai 2020. Encore un outil d'upscaling.

[790] La source de l'image est Context Encoders: Feature Learning by Inpainting, 2016 (12 pages).

Des zones de formes arbitraires peuvent être également remplies avec des GANs[791], illustré *ci-dessous*. La technique a été également développée la même année dans le réseau convolutif **Deep Image Prior**[792].

Certains réseaux génératifs sont plus utiles comme **WESPE** qui améliore les photos prises par smartphone notamment au niveau contraste et basses lumières[793] ou celui du stabilisateur vidéo du Pixel 2 de Google[794].

Une équipe de **Facebook** a créé un GAN servant à modifier les yeux d'une personne dans une photo, parfois mieux qu'avec une retouche photo classique[795]. Cela permet par exemple de corriger les yeux fermés des nombreuses personnes figurant sur une photo de groupe, le cauchemar des photographes ! Mais cela ne fonctionne qu'en basse résolution, comme de nombreuses prouesses à base d'IA de traitement des images.

En 2018, **Nvidia** présentait une solution de ralenti vidéo réalisé à base de GPU et de réseau génératif. C'est impressionnant mais cela doit fonctionner seulement à basse résolution (vidéo). Une autre application consiste à prévoir la suite d'une vidéo[796].

Transfert de style

L'application d'un style à une photo ou une image est une autre application classique des GANs. Elle est souvent montée en épingle comme relevant de la créativité alors qu'il s'agit d'un automatisme déterministe. Ses applications commerciales sont d'ailleurs encore rares.

[791] Voir Globally and Locally Consistent Image Completion, 2017 (14 pages).

[792] Voir Deep Image Prior par Dmitry Ulyanov, Andrea Vedaldi et Victor Lempitsky (10 pages).

[793] Voir WESPE: Weakly Supervised Photo Enhancer for Digital Cameras, d'une équipe de l'ETH Zurich 2017 (10 pages).

[794] Voir Fused Video Stabilization on the Pixel 2 and Pixel 2 XL, 2017 (vidéo).

[795] Voir Eye In-Painting with Exemplar Generative Adversarial Networks, d'une équipe de Facebook, juin 2018 (10 pages).

[796] Voir Generating Videos with Scene Dynamics, 2016.

C'est une application directe des réseaux de neurones convolutifs capables de détecter des features associés à des auto-encodeurs, capables de ré-encoder ces features à partir d'autres bases.

Les cas connus relèvent de l'application du style graphique d'un peintre à une image ou à une photo[797]. C'est plus facile à réaliser avec une image de dessin animé comme *ci-dessus* car la plausibilité du résultat est moins remise en question par l'œil humain[798]. Mais cela fonctionne plutôt bien avec des visages (*ci-dessus*, à droite)[799]. C'est un procédé très mécanique qui n'est pas aussi créatif que l'on pourrait le croire[800] ! L'exemple en bas provient de Li et Wand[801].

À l'envers, on peut transformer un personnage en croquis de dessin animé comme dans l'application mobile **Microsoft Pix** (iOS) lancée en 2017.

Cette méthode s'améliore d'année en année en s'alimentant de données d'entraînement encore plus riches et variées[802].

Cela devient plus impressionnant avec ce GAN qui génère une image plausible à partir d'un simple schéma (*ci-contre*)[803]. A ceci près que les schémas sont eux-mêmes générés à partir de photos, ce qui facilite très probablement le travail du GAN utilisé dans le projet.

Figure 16: Example results of our method on automatically detected edges→shoes, compared to ground truth.

[797] Voir Painting like Van Gogh with Convolutional Neural Networks, novembre 2016.

[798] Et puis aussi, cette tranformation d'un cheval en zèbre dans une vidéo (vidéo).

[799] Voir A Neural Algorithm of Artistic Style par Leon A. Gatys et al, 2015 (16 pages).

[800] Voir Can AI make anyone an artists, septembre 2017. On y trouve aussi la vaste plaisanterie pour gogos de pix2code, une AI qui serait capable de créer un programme à partir d'une simple interface utilisateur, la demonstration étant faite avec une interface comportant deux boutons.

[801] Voir StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, 2016-2017.

[802] Voir Learning to Cartoonize Using White-box Cartoon Representations par Xinrui Wang et al, 2020 (10 pages).

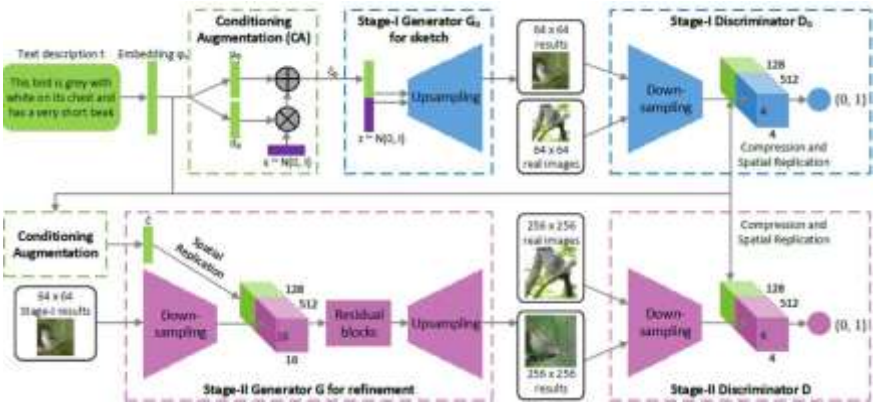
[803] Voir Image-to-image translation with conditional adversarial networks, 2016 (17 pages) et Image-to-Image Demo Interactive Image Translation with pix2pix-tensorflow de Christopher Hesse, février 2017.

Autre variante dans le transfert de style, la transformation d'une image d'extérieur prise en hiver en image d'été chez Nvidia [804]. Là encore, le système a eu besoin d'exploiter une belle base d'entraînement pour générer ce résultat impressionnant.

Le concept du GAN est aussi appliqué au dessin de paysages à main levée avec **GauGAN** (vidéo)[805]. Il suffit de dessiner de manière schématique ce que l'on souhaite obtenir et le GAN se charge du reste et propose des images réalistes. La généralisation d'un tel système nécessite probablement une grosse base d'entraînement pour que cela fonctionne bien au-delà des démonstrations qui sont toujours très sélectives et trompeuses.

Création ex-nihilo

Les autres formes intéressantes de réseaux génératifs sont ceux qui ont la capacité de générer une image plausible à partir d'un descriptif textuel plus ou moins précis. Il faut évidemment disposer d'une très bonne base d'entraînement pour ce faire avec plein d'images taggées avec de nombreux attributs.



Les images générées sont loin d'être parfaites, mais elles trompent facilement le cerveau dans les exemples fournis par les chercheurs[806].

L'exemple *ci-dessous* à gauche est fait de réseaux de neurones multimodaux, associant textes et images [807]! Le réseau génératif de ce type le plus récent est **DALL-E** d'OpenAI, qui un dérivé de GPT-3 exploitant 12 milliards de paramètres couplé à un discriminateur sélectionnant les 32 meilleures images générées par lots de 512 (exemple *ci-dessus* à droite)[808].

[804] Voir dans Unsupervised Image-to-Image Translation Networks, 2018.

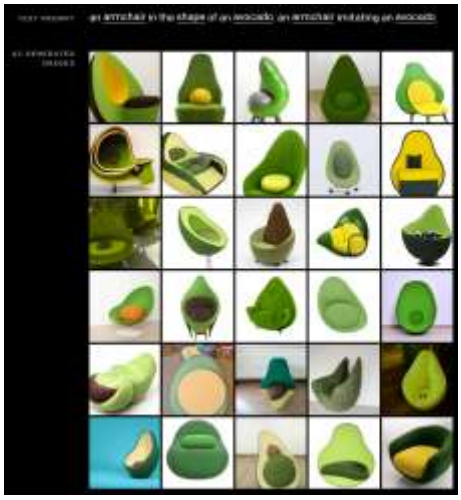
[805] Voir Nvidia AI turns sketches into photorealistic landscapes in seconds, mars 2019 et GauGAN: semantic image synthesis with spatially adaptive normalization, mars 2019 (20 pages).

[806] Les publications scientifiques des réseaux génératifs négligent souvent un point clé : la proportion des images générées qui ne sont pas correctes et que le cerveau humain ne reconnaît pas alors que le discriminateur utilisé dans les GAN les a considérées comme des images plausibles.

[807] Voir StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, 2017 (14 pages).

[808] Voir DALL·E: Creating Images from Text, janvier 2021.

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks par Han Zhang et al, 2017 (14 pages)



OpenAI DALL-E, basé sur GPT-3, janvier 2021

2D en 3D

Différentes techniques de GAN permettent de transformer des objets ou images 2D en représentations volumiques en 3D. C'est encore approximatif comme pour bon nombre de GANs, mais tout de même impressionnant.

C'est le cas de cet exemple qui à partir d'une photo 2D d'une personne est capable de transférer la pose avec une rotation du corps ou bien un changement de tenue (*ci-dessous*)[809].

La méthode **I2L-MeshNet** génère un modèle 3D filaire précis d'une personne prise dans une photo, ou une image extraite d'une vidéo (*ci-dessous*, à gauche).[810].

C'est une amélioration significative par rapport à **DeepPose** de Google, qui date de 2014[811]. Le modèle **PIFuHD** génère un effet voisin (*ci-dessous*, à droite)[812]. La technique a aussi été développée par **Nvidia**[813].

---

[809] Voir Neural Re-Rendering of Humans from a Single Image par Kripasindhu Sarkar et al, 2020 (18 pages).

[810] Voir I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image par Gyeongsik Moon et Kyoung Mu Lee, août 2020.

[811] Voir DeepPose: Human Pose Estimation via Deep Neural Networks par Alexander Toshev et Christian Szegedy, 2014 (9 pages).

[812] Voir PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization par Shunsuke Saito et al, avril 2020 (10 pages) (vidéo).

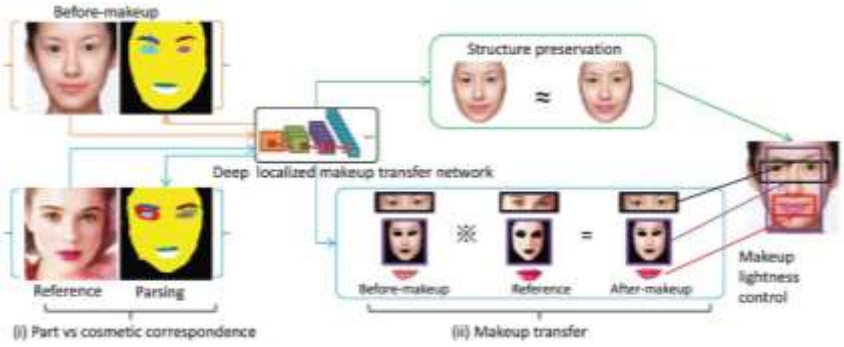
[813] Voir NVIDIA's AI Can Turn 2D Images Into 3D Models par Tyler Lee, novembre 2019.

Visages

Les GANs sont très employés pour modifier, améliorer ou générer des visages.

On trouve ce genre de fonctionnalités dans des applications mobiles comme avec l'application mobile **FaceApp** (Russie) qui peut vous vieillir, vous rajeunir et même changer votre sexe[814]. Elle défrayait la chronique en 2019 parce que ses conditions générales d'utilisation permettaient à la startup d'exploiter les photos récupérées.

L'amélioration de selfies fait aussi partie des projets d'**Adobe** qui utilise son IA maison Sensei, exploitant du deep learning pour améliorer de manière semi-automatique les selfies pris avec des smartphones (vidéo) comme pour corriger les perspectives et divers paramètres de prise de vue a posteriori.



De nombreuses fonctionnalités d'amélioration des portraits issues de Sensei sont maintenant intégrées dans Photoshop et Premiere Elements[815].

Le maquillage virtuel passe par une analyse du visage pour le décomposer en parties auxquelles sont appliquées ensuite divers produits de cosmétique[816], qui peuvent ensuite être bien évidemment commandés en ligne. Nombre de startups du secteur proposent maintenant cela. L'une d'entre elles, **Modiface** (2007, Canada) a été acquise par L'Oréal en mars 2018[817].

Mais une équipe de chercheurs chinois a voulu aller dans le sens inverse, créant une photo d'une femme sans maquillage à partir d'une photo avec maquillage[818]. Cela marche cependant moins bien que l'ajout de maquillage !

La génération d'avatars 3D animés à partir d'une simple photo, utilisant une technique connue de réseau de neurones convolutif génératif. Elle est notamment proposée par la startup américaine **Loom.ai** (2016, $1,35M), créée par des anciens de Dreamworks et LucasFilm (vidéo).

---

[814] Seulement dans la photo !

[815] Voir Adobe announces Photoshop and Premiere Elements 2021 par Jeremy Gray, octobre 2020.

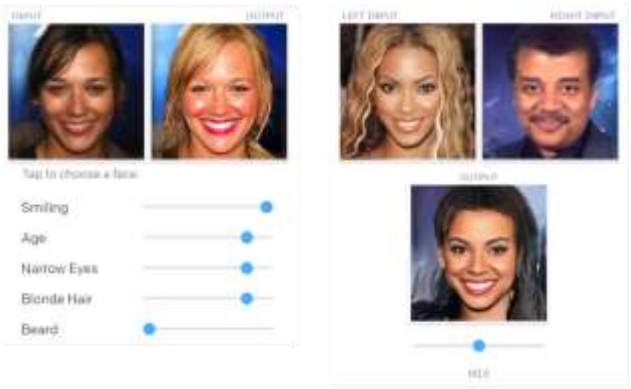
[816] Voir Makeup Like a Superstar: Deep Localized Makeup Transfer Network de Si Liu, Xinyu Ou, Ruihe Qian, WeiWang et Xiaochun Cao, 2016.

[817] D'autres sociétés font cela. Voir AI picks your 'perfect' makeup shade without seeing your face, par Rachel England, juillet 2019.

[818] Voir Anti-Makeup: Learning A Bi-Level Adversarial Network for Makeup-Invariant Face Verification 2018, (8 pages).

Une équipe de Nvidia a pu générer des photos de relativement haute résolution (1024 pixels de côté) d'acteurs qui n'existent pas, grâce à un entraînement progressif du générateur du GAN, avec ajout de couches étape par étape pour double la résolution spatiale[819]. Ce GAN a été entraîné avec le dataset **Celeba**, qui contient des photos de célébrités avec 203 000 photos de 10 177 personnes.

L'un des modèles génératifs les plus récents est **GLOW**[820], créé en juillet 2018 par **OpenAI**. Ce réseau réversible utilise des convolutions de 1x1 pixel. Il permet de générer des images à haute résolution et d'en modifier diverses propriétés comme le sourire, l'âge, la couleur des cheveux et la pilosité. Le modèle permet aussi de mélanger deux visages.



Dans le genre glauque, **DeepFake** peut générer des vidéos de porno avec des personnes qui y sont intégrées à l'insu de leur plein gré[821].

Le buzz a été déclenché avec une démonstration plaquant le visage de l'actrice Gal Gadot sur le corps d'une actrice de porno, habillée pour le besoin du buzz, une fois n'est pas coutume[822]. La menace est telle que le Sénateur du Nebraska, **Ben Sasse**, s'en est même ému en octobre 2018 en avertissant de l'impact potentiel de cette technologie sur les débats politiques aux USA[823].

Une technique voisine est utilisée à des fins moins répréhensibles pour faire danser en vidéo une personne qui ne sait pas danser en imitant une autre personne qui danse[824]. Et Google fait quasiment pareil en imitant vos mouvements avec une saccade de photos de personnes qui sont dans la même position, dans **Move Mirror** (vidéo)[825].

Sont donc lancés des projets divers pour détecter et supprimer les DeepFake. C'est le cas d'un projet américain qui vise à détecter les bizarreries de clignements d'yeux dans les vidéos pour identifier les fake[826]. Jusqu'au jour où les fausses vidéos imiteront bien le clignement des yeux, ce qui ne devrait pas être trop difficile à faire !

Gfycat (2013, USA, $10M) un site de partage GIFs qui détecte les fake on ne sait pas trop comment puisqu'il s'agit d'une startup et pas d'un laboratoire de recherche qui publie ses travaux.

---

[819] Voir Progressive Growing of GANs for Improved Quality, Stability, and Variation, 2017 (26 pages). Voir aussi ces GAN de génération de visages de Nvidia, avec meilleur contrôle des visages générés qui sont des sortes de GAN explicables (vidéo) dans A Style-Based Generator Architecture for Generative Adversarial Networks, par Tero Karras & Al, Nvidia, décembre 2018 (12 pages).

[820] Voir https://blog.openai.com/glow/.

[821] Les vidéos **DerpFake** déclinent la méthode en plaquant diverses personnalités sur des acteurs dans des extraits de films comme pour les James Bond (vidéo), le pire étant, à la fin, avec Rowan Atkinson qui remplace Daniel Craig dans Casino Royale.

[822] Voir AI-Assisted Fake Porn Is Here and We're All Fucked, décembre 2017.

[823] Voir This new technology could send American politics into a tailspin, Ben Sasse dans le New York Times, octobre 2018.

[824] Voir Everybody Dance Now de Berkeley, 2018 (9 pages).

[825] Voir Move Mirror: You move and 80,000 images move with you, 2018.

[826] Voir In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, 2018 (7 pages).

Quant à **Truepic** (2016, USA, $10,5M), ils proposent d'éviter les fake dans les photos ! Comment ? Sans IA. Juste en intégrant un élément de watermark dans les photos générées.

Nous avons aussi un projet en cours de **SRI International** et de l'**Université d'Amsterdam** qui est financé par la **DARPA** dans le cadre du programme **MediFor**, pour Media Forensics. Il vise également à détecter des images et vidéos trafiquées[827].

Des chercheurs russes de **Samsung** sont capables d'animer un visage issu d'une photo en le plaquant sur une vidéo d'un autre visage[828]. Les GANs servent aussi à générer des avatars intégrables ensuite dans des jeux vidéo ou des clips comme cela a été récemment expérimenté en Chine[829].

Un projet brésilien issu d'une Université Catholique financée par le diocèse local rajoute automatiquement des bikinis sur les photos de nues[830].

L'idée est de censurer les images sans que cela se voie trop.

C'est une approche différente de celle de Facebook qui supprime automatiquement les photos contenant des seins, même parfois dans des œuvres d'art bien connues.

Mais ces petits malins créateurs de **DeepNudes** démontrent aussi l'inverse, enlevant le bikini (virtuellement) à quelques femmes pour révéler leur tenue d'Eve, qui n'est heureusement qu'une interpolation[831].

Dans l'application de dating **Bumble**, une IA sert à supprimer les photos de membres qui dévoileraient leur intimité de manière inconvenante[832]. Cela concerne en particulier les hommes qui n'hésitent pas à montrer leurs attributs masculins pour draguer, ce qui constitue une forme de harcèlement pour les femmes qui reçoivent les messages.

La génération de visages avec des GANs est aussi exploitable pour remplacer des visages dans des vidéos, y compris en haute résolution. C'est utilisable dans la production cinématographique comme le présente **Disney**[833].

---

[827] Voir DARPA is funding new tech that can identify manipulated videos and 'deepfakes', de Taylor Hatmaker, avril 2018.

[828] Voir Samsung's AI animates paintings and photos without 3D modeling par Khari Johnson, mai 2019. Des travaux équivalents ont été publiés dans Few-Shot Adversarial Learning of Realistic Neural Talking Head Models par Egor Zakharov & Al, mai 2019 (19 pages).

[829] Voir This Company Is Betting the Future Is Personal AI Avatars, So It Made Me One par Brian Merchant, mai 2019.

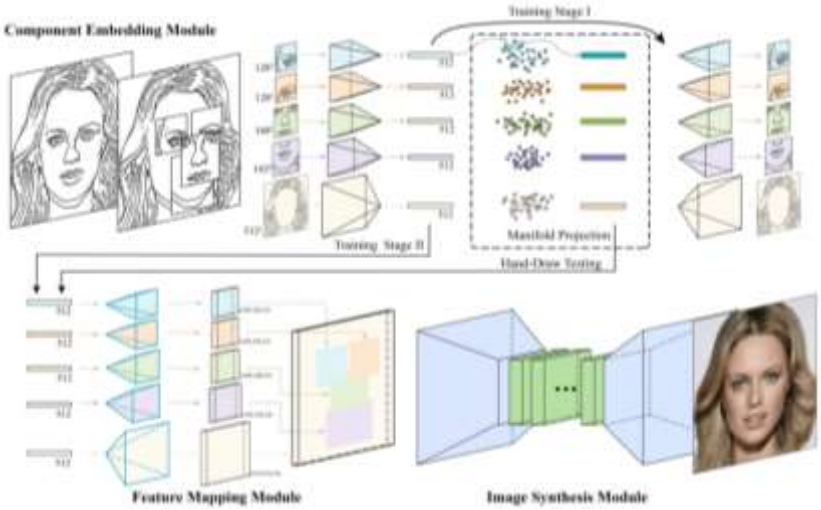
[830] Voir Seamless Nudity Censorship an Image-to-Image Translation Approach based on Adversarial Training, 2018 (8 pages).

[831] Voir 'DeepNude' app to 'undress' women shut down after furor, juin 2019. L'application a été fermée après le tollé généré sur les réseaux sociaux.

[832] Voir Bumble will use AI to detect unwanted nudes par Amrita Khalid, avril 2019.

[833] Voir High-Resolution Neural Face Swapping for Visual Effects, par Jacek Naruniec et al, juin 2020 (vidéo).

Les GANs sont exploitables pour transformer des croquis de visages en visages photo-réalistes. C'est l'objet du projet comme DeepFaceDrawing, qui est issu de chercheurs de Hong-Kong (*ci-contre*)[834].



Rosebud.ai (2019, USA, $1,5M) propose pour sa part sa solution Generative Photos qui permet de se créer une iconographie de visages « politically correct » (couleur, race, genre, etc).

Elle exploite 25 000 photos libres de droit générées par un GAN qui peuvent être mélangées avec des photos proposées. Cela sert à se créer une base de portraits correspondant à ses objectifs de diversité[835].

Peinture

La peinture à base de réseaux génératifs donne lieu à des performances plus ou moins convaincantes selon les cas.

Il y a eu par exemple cette génération de tableau artificiel de Rembrandt réalisée avec l'aide de **Microsoft**[836]. En octobre 2018, le premier tableau réalisé à base d'IA générative représentant un personnage fictif, **Edmond de Belamy**, a été vendu aux enchères chez Christie's pour $432K alors qu'il avait été initialement estimé entre $7K et $10K. De quoi faire jaser. L'auteur est le collectif d'artistes français, **Obvious**, qui regroupe Hugo Caselles-Dupré, Pierre Fautrel et Gauthier Vernier. Leur IA à base de GAN a été entraînée avec 15 000 portraits peints entre le 14e et le 20e siècle. Une fois encore, il faut rappeler que l'IA n'a pas réalisé toute seule cette peinture. Ce sont des peintres outillés par l'IA qui l'ont créée [837] !

Il y a encore plus fort avec ce GAN Timelapse créé par le laboratoire CSAIL du MIT qui est capable de reconstituer les étapes de création d'une peinture à partir du résultat final, exploitant comme il se doit des données d'entraînement avec le cycle de production de nombreux tableaux[838].

---

[834] Voir DeepFaceDrawing Generates Photorealistic Portraits from Freehand Sketches, juin 2020.

[835] Voir Generative Photos : 25 000 photos libres de droit avec des visages générés par une IA par Éléonore Lefaix, novembre 2019.

[836] Voir The Next Rembrandt.

[837] Voir Is artificial intelligence set to become art's next medium?, Christie's, octobre 2018. A noter que le collectif Obvious a utilisé une partie du code provenant d'une autre artiste, Robbie Barrat, qui l'avait partagé sous une licence open source (certes, lisible, utilisable et modifiable par tous).

[838] Voir Using AI to recreate how artists painted their masterpieces, juin 2020 (vidéo).

Langage

Le traitement du langage est le second plus grand domaine d'applications de l'IA avec celui de l'image. Il comprend de nombreuses fonctions et notamment la reconnaissance de la parole, les robots conversationnels, la traduction automatique, l'extraction de données, la création de résumés et la génération de textes.

Ces outils couvrent tout le spectre qui va de la compréhension du langage à son interprétation, son exploitation puis à la création de textes ou de paroles. Il comprend aussi les outils et méthodes de représentation des connaissances.

Ce domaine exploite surtout le deep learning et les réseaux récurrents et à mémoire. Ce champ de l'IA est cependant un peu moins mature que celui de l'image. Autant, par exemple, peut-on dire qu'une IA de diagnostic dans l'imagerie médicale équivaut à celle d'un spécialiste, autant un chatbot est encore loin de passer avec succès le test de Turing et d'arriver à se faire passer pour un Humain[839]. Ou tout simplement, à conduire une discussion cohérente de bout en bout pour des demandes élémentaires, même dans l'environnement calme de son logement.

Le deep learning appliqué au langage est aussi probabiliste que celui qui est appliqué aux images. Celui-ci a permis de générer d'énormes progrès dans tous les domaines du traitement du langage, en gros, entre 2012 et 2017[840].

Ces avancées du deep learning sont parfois remises en cause par des chercheurs et entrepreneurs qui trouvent que ces approches probabilistes ont des limites. Ils remettent au gout du jour des méthodes qui réinjectent un peu de symbolisme dans les procédés employés[841].

---

[839] Un agent conversationnel est censé avoir passé le test de Turing en 2014. Voir https://en.wikipedia.org/wiki/Eugene_Goostman. Mais en imitant un adolescent de 13 ans dans une discussion assez limitée. Donc, le véritable test de Turing n'est pas encore véritablement passé. Même Ray Kurzweil considère qu'il faudra patienter jusqu'à 2030 pour y arriver. Il existe un concours international visant à récompenser l'agent conversationnel le plus proche d'un humain, le Loebner Prize dont la première édition avait lieu en 2006 et la dernière en 2019. Les quatre dernières éditions ont été remportées par le programme Mitsuku de la société Pandorabots (2008, USA).

[840] Voir cet excellent historique des avancées dans le traitement du langage : A Review of the Neural History of Natural Language Processing, de Sebastian Ruder, octobre 2018.

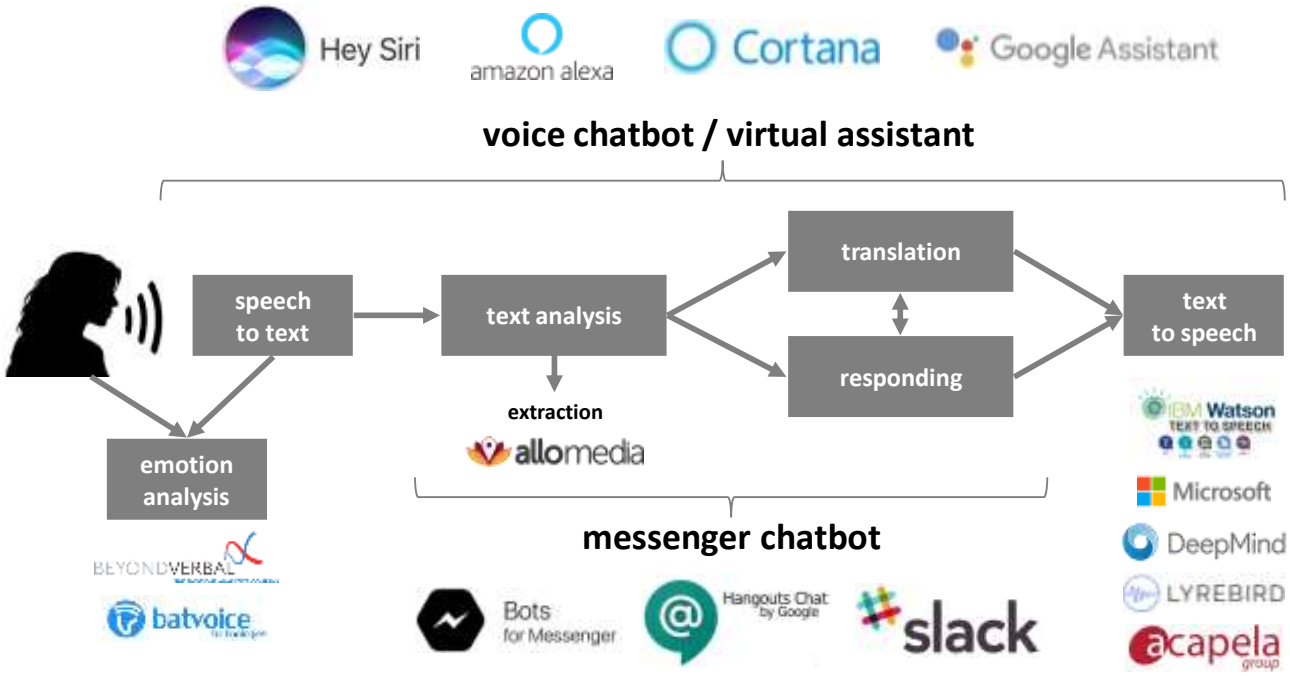
[841] Voir par exemple le champ des Memory Networks, 2014, qui gère la mémoire long terme pour les applications de questions/réponses, et les Logic Tensor Networks proposés en 2016 qui combinent les réseaux de neurones avec de la logique symbolique et de la logique floue. What are "Logic Tensor Networks"? de Lucas Bechberger, novembre 2017, Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge, 2016 (12 pages) et la présentation associée Learning and Reasoning in Logic Tensor Networks, mai 2017 (38 slides).

Ce que l'on retrouve aussi bien chez Google avec ses **Universal Transformers** ou des startups telles que le Français **Golem.ai**, spécialiste des chatbots.

Le champ du traitement du langage et de la représentation des connaissances est très riche et en évolution constante. Il est très difficile à suivre pour les néophytes. Les concepts se renouvellent sans cesse[842].

Ainsi en va-t-il par exemple des **Knowledge Graph Embeddings** et des **Poincaré Embedding** qui sont notamment utilisés par Facebook et qui visent à représenter efficacement les liens entre expressions textuelles[843].

Dans le schéma *ci-dessous*, je présente de manière synthétique une partie des éléments qui vont suivre et qui positionnent le rôle des chatbots textuels, de la traduction automatique, des assistants vocaux, des systèmes de speech to text et d'analyse des émotions ainsi que des générateurs de parole synthétique. Tous sont potentiellement reliés les uns aux autres. Ce sont les assistants vocaux qui intègrent le plus grand nombre de briques d'IA et sont les plus difficiles à mettre au point. Il n'est ainsi pas étonnant qu'ils soient surtout maîtrisés par les GAFAMI.



Le traitement du langage a aussi des usages originaux comme faire survivre des langues indigènes rares, comme au Canada ou à Hawaï[844].

---

[842] Voir Recent Trends in Deep Learning Based Natural Language Processing de Tom Youngy, Devamanyu Hazarikaz, Soujanya Poria et Erik Cambria qui fait un bon état des lieux, février 2018 (24 pages). Les principales techniques de deep learning citées sont le word embeddings et Word2vec (vecteurs de mots qui contiennent les mots apparaissant dans le contexte de mots donnés), les réseaux convolutifs utilisés pour la modélisation de phrases pouvant notamment servir à la production de résumés, les réseaux de neurones récurrents (RNN, LSTM) qui sont notamment utilisés dans la traduction, les mécanismes de gestion de l'attention qui associent des convnets et des LSTM, les modèles génératifs et les réseaux à mémoire. Pour les plus courageux, voici Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, de Stanford, 2018 (588 pages).

[843] Voir Knowledge Graph Embeddings: Recent Advances de William Wang, 2018 (21 slides) et Implementing Poincaré Embeddings, de Jayant Jain, décembre 2017.

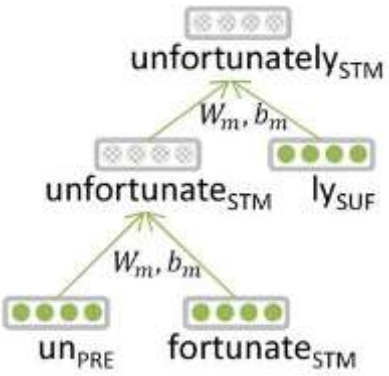
[844] Voir Turning To AI To Save Endangered Languages par Stephen Ibaraki, 2018.

Reconnaissance de la parole

La reconnaissance de la parole est la première étape du dialogue naturel entre un humain et une machine. Elle vise à transformer la voix en texte lisible par un humain et ensuite, traitable par la machine comme le serait un texte que l'on aurait tapé au clavier.

Elle s'appuyait au départ sur des modèles cachés de Markov et des réseaux probabilistes bayésiens. Ces algorithmes généraient des taux d'erreur assez élevés[845].

La complexité de la reconnaissance de la parole provient des variantes dans l'expression qui dépendent du contexte et aussi des ambiguïtés du langage. Les progrès récents sont dus à l'adoption du deep learning qui s'est avéré bien plus efficace que les anciennes méthodes[846]. Les systèmes de reconnaissance de la parole utilisent souvent des réseaux convolutifs pour reconnaître les phonèmes à partir du signal audio qui est analysé sous la forme d'un spectre de fréquences dans le temps[847]. Ils les assemblent ensuite avec des algorithmes qui permettent d'identifier la morphologie des mots, par assemblage de phonèmes.



S'en suit, en général avec des réseaux récurrents, l'assemblage de mots dans des phrases ou locutions. Une fois transformés en mots, ceux-ci sont transformés en vecteurs avec autant de bits que de mots dans le dictionnaire, et un seul bit à 1 pour l'indice du mot dans le dictionnaire. Y sont ajoutés ensuite les poids des mots avoisinants dans des phrases connues avec une valeur correspondant à leur probabilité d'apparition autour d'un mot (le mot de référence a donc une probabilité de 1 et les autres, une probabilité inférieure à 1).

Ces vecteurs sont ensuite compressés, (Dense Vector) comme comme via la technique dite Word2Vec, créée par Google en 2013. Cela sert à gagner de la place en ne conservant que les indices et poids des mots pertinents. Ces vecteurs sont utilisés de diverses manières. On peut même réaliser des opérations de logique avec les mots ainsi modélisés (*ci-contre*). Paris moins la France plus l'Italie devient ainsi Rome !

Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

Le reste est une tambouille mathématique de vecteurs qui dépend des nombreuses méthodes utilisées ! Les IA de reconnaissance de la parole manipulent des symboles mathématiques et n'ont qu'une vision purement statistique du langage. Elle n'est pas du tout symbolique. La machine ne comprend pas ce qu'elle interprète. Les progrès de la reconnaissance de la parole se sont accélérés depuis l'utilisation intensive du deep learning avec de nombreuses couches de neurones, jusqu'à 1000 !

---

[845] Pour en savoir plus, voir cet historique de la recherche en reconnaissance de la parole Survey of Technical Progress in Speech Recognition by Machine over Few Years of Research, 2015 (7 pages), qui commence d'ailleurs à dater. Ce sujet intègre de nombreuses branches du savoir issu de plusieurs décennies de recherches dans l'IA.

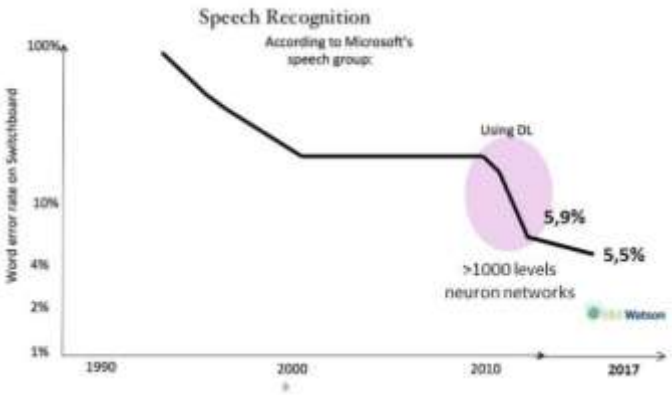
[846] Les premiers progrès en deep learning sont arrivés avec Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, de George Dahl, Dong Yu, Li Deng et Alex Acero, 2010 (13 pages), quis'appuyait encore sur des modèles de Markov. Ces chercheurs de Microsoft Research ont utilisé cette méthode pour faire descendre le taux d'erreurs de la reconnaissance de la parole de 23% à 13% en 2012.

[847] A partir de 2011, on a pu commencer à se passer des transformées de Fourier pour convertir le signal audio en spectre de fréquences, en analysant l'onde audio directement dans des réseaux de neurones. Voir Deep Learning for Speech/Language Processing - machine learning & signal processing perspectives de Li Deng, 2015, slide 123 (200 slides).

Le taux d'erreur est maintenant inférieur à celui de la compréhension humaine, que ce soit chez **Microsoft** et **IBM**. Mais il semble qu'il ne s'agit que de taux de reconnaissance de mots car on peut bien observer au quotidien que les assistants vocaux comprennent à peine deux tiers de nos paroles, comme une grand-mère malentendante.

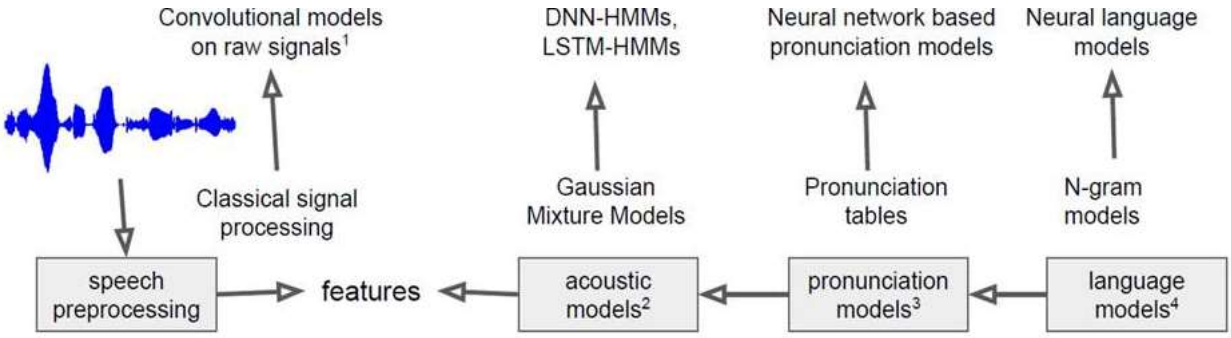
Les solutions de reconnaissance vocale à base de deep learning n'ont plus besoin, en théorie, d'être entraînées avec la voix de l'utilisateur.

Les modèles de ces réseaux sont créés avec des bases de tests comme pour TIMIT pour l'Américain, qui comprend les mots de 630 locuteurs différents. Des techniques d'apprentissage par renforcement existent cependant qui affinent la qualité des modèles utilisés.



La reconnaissance de la parole peut être réalisée localement ou sur des serveurs. Avec l'augmentation de la puissance des processeurs embarqués dans les mobiles et même dans certains objets connectés, il est de moins en moins nécessaire de faire un aller et retour avec des serveurs dans le cloud.

Lorsqu'un aller et retour est nécessaire, on voit tout l'intérêt de la 4G avec son débit comme et sa faible latence pour les allers et retours avec les serveurs. Et cela sera encore mieux avec la 5G.



1. Jaitly, Navdeep, and Geoffrey Hinton. "Learning a better representation of speech soundwaves using restricted boltzmann machines." *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011.
2. Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal Processing Magazine* 29.6 (2012): 82-97.
3. Rao, Kanishka, et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.
4. Mikolov, Tomas, et al. "Recurrent neural network based language model." *Interspeech*. Vol. 2. 2010.

On est encore loin de la solution parfaite [848], notamment parce que les logiciels manquent d'informations sur le contexte des conversations[849]. Le taux de fiabilité n'est jamais de 100%. Il ne l'est d'ailleurs jamais pour l'Homme également !

Le taux d'erreur de la reconnaissance est toujours plus élevé dans d'autres langues comme le chinois sans compter les langues rares pour lesquelles les bases d'entraînement sont moins grandes que pour les grandes langues parlées dans le monde.

D'où l'intérêt de la publication en open source de la solution **Deep Speech 2** de **Baidu** qui fonctionne en anglais et en chinois[850].

[848] Voir Will computers ever truly understand what we're saying?, janvier 2016.

[849] Voir aussi Why our crazy smart AI still sucks in transcribing speech paru dans Wired en avril 2016.

Le taux d'erreur est particulièrement élevé dans un environnement bruyant, comme dans la rue, dans un endroit où il y a du monde et même dans sa voiture. Des techniques de captation du son et d'élimination du bruit ambiant permettent de traiter en partie ce problème.

Certaines portent sur l'analyse spectrale et le filtrage de fréquences. D'autres utilisent la captation stéréophonique pour séparer le bruit proche (différencié) du bruit lointain (qui l'est moins). J'avais même vu la start-up israélienne **VocalZoom** (2010, Israël, $12,7M) au CES 2015 qui utilisait un laser pour capter les vibrations des lèvres. Il faut juste trouver où placer le laser, ce qui est plus facile sur des installations fixes que mobiles.

Une autre manière d'améliorer la reconnaissance de la parole est de faire du « multimodal », à savoir capter plusieurs signaux en même temps comme la voix et une vidéo du locuteur. C'est ce qu'ont réalisé des chercheurs de Google[851]. On y voit deux anglophones parler en même temps dans un environnement bruyant. Le système est alors capable d'isoler une à une les deux voix en éliminant l'autre personne et le bruit ambiant. C'est ce que fait notre cerveau lorsque l'on suit plusieurs personnes à la fois dans un dîner bruyant !

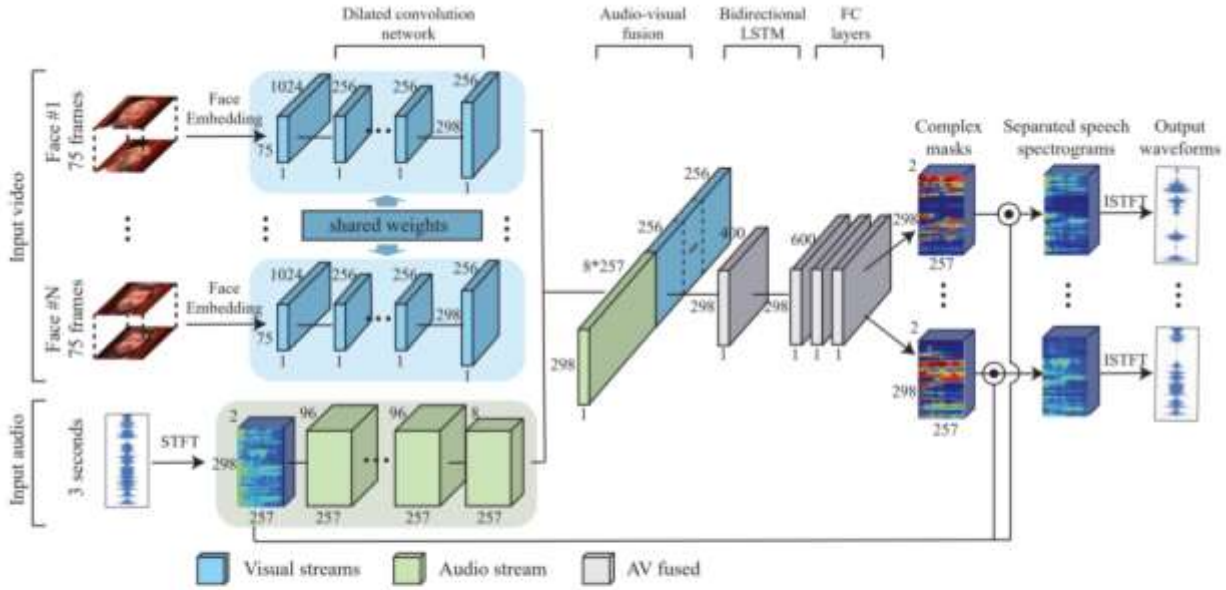


Fig. 4. **Our model's multi-stream neural network-based architecture:** The visual streams take as input thumbnails of detected faces in each frame in the video, and the audio stream takes as input the video's soundtrack, containing a mixture of speech and background noise. The visual streams extract face embeddings for each thumbnail using a pretrained face recognition model, then learn a visual feature using a dilated convolutional NN. The audio stream first computes the STFT of the input signal to obtain a spectrogram, and then learns an audio representation using a similar dilated convolutional NN. A joint, audio-visual representation is then created by concatenating the learned visual and audio features, and is subsequently further processed using a bidirectional LSTM and three fully connected layers. The network outputs a complex spectrogram mask for each speaker, which is multiplied by the noisy input, and converted back to waveforms to obtain an isolated speech signal for each speaker.

Il existe des besoins particuliers dans la reconnaissance vocale. Par exemple, pour reconnaître la voix des enfants. C'est la spécialité de **Soapbox Labs** (2013, Irlande, $11,8M) qui analyse la voix des enfants dans la classe ou les cours de récréation, mais avec une API dans le cloud.

Se pose aussi la question de la reconnaissance de la parole de personnes atteintes de handicaps liés par exemples à des maladies neurodégénératives (vidéo).

---

[850] Voir Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, décembre 2015. Ce système fonctionne avec un réseau de neurones de 11 couches : 3 couches de convolution pour la reconnaissance des phonèmes, sept couches de réseaux de neurones récurrents pour la détection des mots, puis une couche de connexion (« fully connected layer »). En mandarin, il obtient un taux de reconnaissance supérieur à l'homme pour des phrases courtes. Il a été entraîné avec 12 000 heures de conversations. Les versions les plus récentes ont été entraînées avec plus de 100 000 heures de conversations en environnement bruyant.

[851] Dans Looking to Listen at the Cocktail Party Speaker-Independent Audio-Visual Model for Speech Separation 2018 (11 pages) et avec des explications en langage plus naturel dans Looking to Listen: Audio-Visual Speech Separation, avril 2018. La vidéo associée est très … parlante !

Google a présenté des travaux de recherche dans ce domaine lors de la conférence Google I/O en mai 2019[852].

Le traitement de la parole contient un sous-domaine relativement récent : la détection des émotions dans la parole. C'est l'offre de diverses startups comme **BatVoice** (2015, France) qui se propose ainsi de capter les émotions des clients appelant un call center et d'évaluer l'efficacité des agents qui y répondent et savent traiter le stress des clients. C'est aussi l'offre d'une autre startup, **BeyondVerbal** (2012, Israël, $10M) qui commercialise de la propriété intellectuelle issue de longues années de recherche dans le domaine. Ils cherchent à détecter des pathologies neurodégénératives avec l'analyse de la parole. Des chercheurs du MIT s'en servent pour détecter des dépressions lors d'entretiens avec des patients[853]. Mais cela reste encore très expérimental. On pourrait même déterminer l'âge d'une personne à sa voix, une solution développée par **Nuance**[854] ! Elle servirait à détecter les seniors pour bien les traiter et à détecter certaines formes de fraude.

Il faut se garder de prendre ces solutions à la lettre. Elles sont très approximatives car elles manquent souvent d'éléments de contexte pour bien interpréter les émotions des locuteurs. Les signes extérieurs de ces émotions ne sont pas suffisants pour comprendre les émotions réelles des gens.

Cogito (2007, USA, $117,5M) analyse les appels dans les centres d'appels aider en temps réel les conseillers en ligne. C'est une spin-off du MIT Media Lab qui exploite les sciences comportementales. Ils comparent les caractéristiques des conversations à un historique en analysant la tonalité, le volume, les pauses et la vitesse des discussions. Le système est censé améliorer de 20% la satisfaction des clients. Dans la même veine, les startups françaises **Natural Talk** (2016, France) et **Cognitive Matchbox** (2016, France) proposent chacune une solution de routage d'appels optimisée aux centres d'appels qui analysent la personnalité et les émotions des clients pour les orienter vers le meilleur agent.

---

[852] Voir Google details AI work behind Project Euphonia's more inclusive speech recognition par Devin Coldewey, août 2019, qui fait reference à Project Euphonia's Personalized Speech Recognition for Non-Standard Speech par Joel Shor et Dotan Emanuel, de Google Research à Tel Aviv, août 2019.

[853] Voir MIT Develops AI That Can Detect Depression From Your Conversations, février 2018. D'autres utilisent les mouvements du visage pour ce genre de détection, comme le projet de recherche Sensei / Multisence de l'USC (University of Southern California). Voir leur vidéo et leur brochure.

[854] Voir This AI can guess your age from the sound of your voice par Jared Newman, juin 2020.

Elles exploitent les APIs d'IBM Watson dédiées au traitement du langage naturel comme Personality Insights, Natural Language Understanding, Tone Analyzer, Document conversion, Twitter Insight et Natural Language Classifier.

AlloMedia (2011, France, $12,3M) utilise la reconnaissance de la parole pour extraire des informations structurées et semi-structurées des dialogues avec les clients dans les centres d'appels, pour alimenter leurs bases de CRM et améliorer la transformation des leads.

C'est ce que proposent également **MonkeyLearn** (2014, USA, $1,2M) ainsi que **Dialpad** (2011, USA, $120M).

La reconnaissance de la parole ne permet bien entendu pas de créer une solution complète. Il faut lui ajouter un système qui comprend le sens des questions et qui y répond ! Il doit exploiter une base de connaissance, des arbres de décision et un convertisseur de texte en langage parlé (text to speech). C'est ce que l'on trouve dans les assistants personnels ou chatbots vocaux selon les appellations, que nous verrons plus loin.

A noter que les réseaux de neurones permettent aussi de classifier les cris d'animaux comme les chats[855] ! On aura tout vu ! Ce sont des modèles probablement extensibles à un grand nombre d'animaux. Par exemple les dauphins ! Vous ne croyez pas si bien dire ! Plusieurs réseaux de neurones convolutifs ont été créés pour classifier les cris des dauphins (2016[856]) et pour déterminer leur espèce (2018[857]). Et cela s'applique aussi aux baleines[858] !

Synthèse vocale

A l'autre bout des assistants personnels se trouvent des systèmes de génération de parole synthétique. L'objectif est de rendre les voix artificielles les plus réalistes possibles, ce qui est assez difficile à réaliser. Jusqu'en 2018, on continuait à reconnaître les voix synthétiques, même avec les meilleurs outils de synthèse vocale. Depuis 2019, la différence commence à sérieusement s'estomper au point de poser des questions d'éthique, liées à l'impossibilité émergente de distinguer une machine d'un humain. D'autres questions d'éthique se posent lorsque par exemple les voix féminines sont surreprésentées dans les agents conversationnels, renforçant le stéréotype « de la servante ». L'ONU s'en est récemment inquiétée[859].

Le text-to-speech est une technique complexe, peut-être pas autant que le speech-to-text, mais elle repose aussi sur l'exploitation de réseaux de neurones récurrents, histoire de savoir comment juxtaposer les phonèmes les uns aux autres en fonction du contenu à lire.

Google a une excellente solution dans le domaine tout comme **Amazon** avec Polly. Ces solutions sont paramétrables pour spécifier le rythme de la génération, l'intonation, et le style de voix.

Microsoft mettait à jour sa technique de text-to-speech à base de deep learning avec Neural TTS en septembre 2018[860]. La qualité des voix générées est de très bon niveau, quasiment impossible à distinguer de voix naturelles. Dans une itération de 2019, le speech-to-text devenait plus facile à entraîner avec moins de données[861].

---

[855] Voir Domestic Cat Sound Classification using Transfer Learning, juin 2018 (7 pages).

[856] Voir Classifying dolphin whistles using convolutional neural networks par Genevieve Flaspohler, 2016 (15 slides).

[857] Voir A convolution neural network for dolphin species identification using echolocation clicks signal, septembre 2018 (4 pages).

[858] Voir Deep Machine Learning Techniques for the Detection and Classifcation of Sperm Whale Bioacoustics, Peter C. Bermant, 2019 (10 pages).

[859] Voir Using female voices for AI assistants hurts women, UN report says par Shelby Brown, mai 2019.

[860] Voir Microsoft's new neural text-to-speech service helps machines speak like people, septembre 2018.

[861] Voir Microsoft Develops AI-Based Text-To-Speech That Required Very Little Training par Tyler Lee, mai 2019.

Des startups abordent aussi sur ce marché comme la canadienne **Lyrebird** (2017, Canada, $120K) qui propose sa solution aux développeurs sous forme d'API en cloud, exploitant des serveurs à base de GPU Nvidia. Elle permet de copier la voix d'une personne à partir d'un court échantillon censé être d'une minute et de contrôler l'émotion dans l'intonation[862]. La startup était acquise en 2019 par **Descript** (2017, USA, $20M) qui permet l'édition de vidéos et podcast via leur transcript textuel.

Acapela Group (1997, Belgique) propose aussi ses briques logicielles de *text to speech* qui sont notamment dédiées à des solutions d'accessibilité.

Dans le même ordre d'idée, l'expérience **JFK Unsilenced** de 2018 faisait lire le discours que JFK devait donner après en début d'après-midi je jour de son assassinat à 12h30, au Dallas Trade Mart[863]. La voix générée est saisissante de réalisme.

La prouesse est cependant techniquement inférieure à celle de Lyrebird avec Barack Obama car elle ne comprend pas la vidéo. Elle est par contre symboliquement beaucoup plus impactante.

L'expérience a été renouvelée en 2019 par le MIT avec la production du discours que Richard Nixon aurait eu à prononcer si la mission Apollo 11 avait échoué[864].

Une équipe de chercheurs de **Baidu** a créé en 2018 une solution de synthèse vocale utilisant la voix d'une personne avec un échantillon très réduit de quelques secondes par rapport à l'habitude[865]. **Facebook** fait de même[866].

Et avec **SING**, Facebook propose une méthode de génération de notes d'instruments de musique plus efficace que WaveNet de Google[867].

L'entreprise de logiciels spécialisée dans le traitement du langage **iFlyTek** (1999, Chine) arrivait de son côté à faire parler Donald Trump en mandarin fin en 2017 (vidéo). On ne peut pas facilement vérifier que c'est plausible ! C'est un concurrent de l'Américain **Nuance** qui a aussi une offre dans le domaine de l'assistance des juges dans les tribunaux.

La difficulté de la génération de paroles est son réalisme. **Google Tacotron 2** qui s'appuie sur **WaveNet** de DeepMind et Tensorflow est exploité dans Google Assistant. Il commence à atteindre un niveau de réalisme qui rend difficile la distinction entre voix artificielle et voix humaine[868].

---

[862] Voir leurs démonstrations avec les voix de Donald Trump et Barack Obama : https://lyrebird.ai/vocal-avatar. Une vidéo synthétique de Barack Obama a été produite plus tard par l'Université de Washington et le résultat est encore meilleur : Fake Obama created using AI video tool - BBC News.

[863] Voir https://rothco.ie/work/jfk-unsilenced/. La performance a reçu le Grand Prix for Creative Data du Cannes Lions International en juin 2018. Voir AI-Driven JFK Unsilenced Triumphs in Creative Data at Cannes de Alexandre Jardine, juin 2018.

[864] Voir President Nixon Never Actually Gave This Apollo 11 Disaster Speech. MIT Brought It To Life To Illustrate Power Of Deepfakes par Bob Shaffer, novembre 2019.

[865] Dans Neural Voice Cloning with a Few Samples, 2018 (17 pages) et les explications dans Neural Voice Cloning with a Few Samples, février 2018

[866] Voir Facebook's AI system can speak with Bill Gates's voice, juin 2019.

[867] Voir SING: Symbol-to-Instrument Neural Generator par Alexandre Défossez (FAIR et PSL) & Al, octobre 2018 (11 pages).

[868] Voir la page https://google.github.io/tacotron/publications/tacotron2/index.html et les comparaisons à la fin de la page entre voix humaine et voix artificielle.

Google propose des imitations de voix de stars comme John Legend. Tout ceci pourrait servir à la production d'audiobooks reprenant la voix de leurs auteurs sans passer par un enregistrement[869].

Agents vocaux

La reconnaissance de la parole est maintenant intégrée dans un nombre croissant de solutions grand public.

Le marché des agents conversationnels vocaux ou textuels est dominé par de grands acteurs américains (**Google** Assistant, **Amazon** Alexa, **Apple** Siri, **Microsoft** Cortana, **Samsung** Bixby qui est probablement originaire de **Viv Labs**[870] sans compter les équivalents chinois comme **Baidu** avec son DuerOS[871]. Ils sont activés par ce que l'on dénomme des « Wake-up Word » qui déclenchent la reconnaissance vocale.

Leurs solutions sont disponibles à la fois dans leurs propres services comme Amazon Echo ou l'iPhone pour Siri mais également disponibles sous forme d'API en cloud exploitables par les développeurs d'applications et de solutions métiers.

Enfin, ils sont intégrés dans des enceintes à commande vocale comme chez Google et son Home, Apple avec son HomePod[872], les Amazon Echo et plein d'autres copycats originaires d'Asie. Ces solutions vont d'ailleurs réduire l'intérêt pour certains usages de faire appel à des télécommandes traditionnelles voire des boutons.

Ce sont des plateformes qui proposent un SDK et l'accueil d'applications spécifiques. Est-ce que pour autant la commande vocale est destinée à remplacer toutes les autres interfaces homme/machine, dont le clavier. Comme l'a montré la lente histoire des interfaces hommes/machine, le clavier et l'écran ne sont pas près de laisser la place à la commande vocale[873]. L'une des raisons est que, si l'on peut parler plus vite que l'on écrit, la lecture est plus rapide que l'écoute. Et l'écrit reste plus discret que la parole lorsque l'on n'est pas seul !

On aura donc besoin pour longtemps d'interfaces utilisateurs multimodales, au même titre que le tactile et la souris se complètent bien sur un ordinateur (sous Windows...) à écran tactile.

---

[869] Voir Artificial Intelligence Can Now Copy Your Voice: What Does That Mean For Humans? par Bernard Marr, mai 2019.

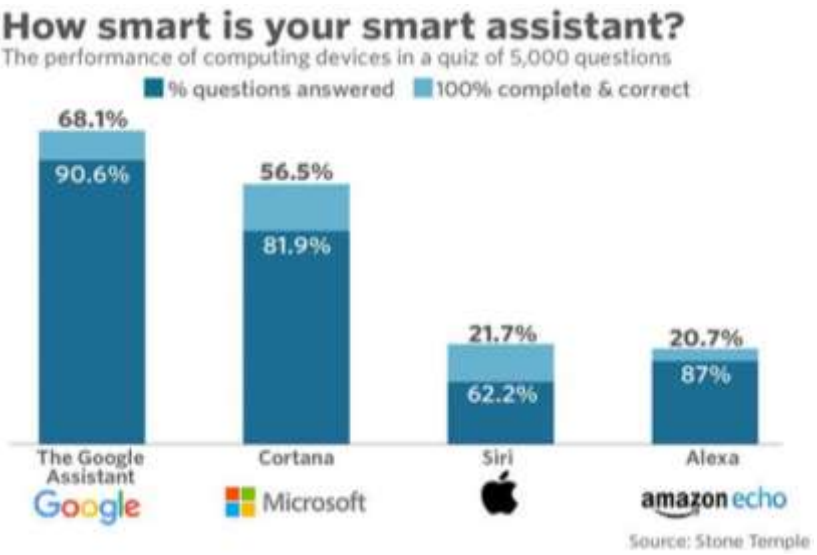
[870] Viv, des créateurs de Siri, est un agent conversationnel capable de répondre à des questions complexes, bien au-delà de ce que peuvent faire Apple Siri et Google. La solution exploite la notion de génération dynamique de programme. Après analyse de la question, un programme complexe est généré en moins d'une seconde qui va la traiter. Viv a été présenté lors de TechCrunch Disrupt à New York (vidéo). Viv Labs (2012, $30M) a été acquis par Samsung pour $215M en 2016.

[871] Voir Baidu Enters the AI Assistant Fray With DuerOS, août 2017.

[872] Voir I tried out Apple's new HomePod features. Here's what I learned, mai 2018 qui décrit les fonctionnalités limitées de Siri. L'Apple HomePod est disponible en France et en français depuis juin 2018.

[873] Voir The Future Is Multimodal: Why Voice Alone Will Never Be the Answer par Tobias Goebel, février 2020.

Ces différents assistants se distinguent dans leur capacité à répondre à des questions diverses. Un benchmark de 2017 mettait les assistants de Google et Microsoft devant ceux d'Apple et Amazon (*ci-contre*[874]). Il se trouve que seuls IBM, Microsoft et Google entretiennent une véritable recherche fondamentale sur le sujet, mais ce n'est peut-être pas la seule explication. Le retard technique d'Amazon explique peut-être l'annonce fin août 2016 d'un partenariat entre Microsoft et Amazon.



Ils font en sorte que Cortana puisse dialoguer avec Alexa[875] et réciproquement, rendant ainsi compatibles, via la voix, leurs bibliothèques de services compatibles. On peut donc dire « Alexa demande à Cortana de demander à trucmuch de faire ceci cela ». C'est encore un peu lourd comme forme d'intégration.

Amazon Alexa est maintenant capable de parler en ajoutant de l'émotion dans sa voix et de capter les émotions de son utilisateur[876]. Cela reste assez mécanique mais peut simuler l'ennui ou l'excitation dans le propos.

Les progrès apparents sont constants avec les agents vocaux. Google a marqué les esprits avec sa démonstration de **Google Duplex** lors de la conférence Google I/O, un assistant vocal qui prend des rendez-vous téléphoniques à votre place[877].

C'est bluffant mais rien ne prouve que cela fonctionne dans une grande diversité de scénarios. Qui plus est, il semblerait que dans certains cas, les rendez-vous soient réalisés par des humains, même pour la démonstration de 2018[878]. La solution est cependant déployée depuis 2019 pour faire des réservations diverses comme celles de restaurants[879]. Microsoft fait de même, mais seulement en chinois[880], et l'efficacité est plus difficile à vérifier pour ce qui nous concerne !

---

[874] Voir Alexa and Cortana Will Talk to Each Other Say Amazon and Microsoft, dans Voicebot.ai, août 2017.

[875] Vous pouvez passer un bon moment avec ces deux parodies d'Amazon Alexa par l'émission Saturday Night Live : Amazon Echo – SNL et Amazon Echo Commercial Parody.

[876] Voir Use New Alexa Emotions and Speaking Styles to Create a More Natural and Intuitive Voice Experience par Catherine Gao, novembre 2020.

[877] Voir Did Google Duplex AI demonstration just pass the Turing test, mai 2018 et la démonstration en vidéo.

[878] Voir 25% des appels Google Duplex réalisés par des humains, mai 2019.

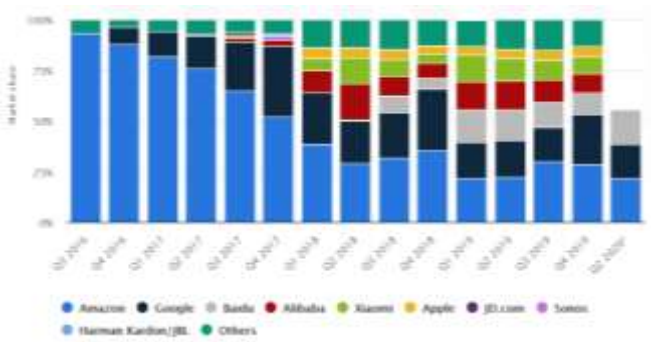
[879] Voir AI-powered booking service Google Duplex rolls out to iOS & Android 5.0+ devices par Sarah Perez, avril 2019.

[880] Voir Microsoft's AI Bot Can Make Phone Calls To Humans As Well, mai 2018, mais c'est en Chine et en Chinois !

La fonction est intégrée dans les smartphones Pixel 3 de Google depuis fin 2018. Dans les progrès moins spectaculaires, Google Assistant n'a plus besoin d'être relancé par « OK Google » pour une question qui suit une autre question[881]. C'est un point clé des agents vocaux : savoir suivre une conversation dans la durée et en préserver les données de contexte.

L'agent vocal peut aussi s'habiller d'un avatar réaliste. C'est le propos d'un projet à venir, **Autodesk Virtual Agent** (AVA) qui est un avatar avec un visage et une voix très réalistes, dont les émotions sont générées par la startup **Soul Machines** (Nouvelle-Zélande, $7,5M), créateurs du Virtual Nervous System (VNS)[882] (vidéo). C'était aussi l'intention du projet Neon présenté par Samsung au CES de janvier 2020. Si les avatars photoréalistes créés à partir de la capture 3D de personnes réelles, la partie vocale était au niveau des agents actuels : décevante. Mais cela s'améliorera sans doute[883].

Les parts de marché des assistants vocaux sont très volatiles depuis trois ans, surtout côté enceintes connectées (*ci-contre*, source Statista). Le marché est dominé par Amazon avec ses Echo, les premiers arrivés aux USA. Google est monté en puissance sur le marché des smart speakers en 2018 puis a décliné en 2019[884]. Sachant que tout cela peut varier d'un pays ou d'une géographie à l'autre en fontion de paramètres marketing divers.



Une nouvelle discipline a fait son apparition : la **VUI** qui est aux interfaces vocales ce que la GUI est aux interfaces graphiques. La Vocal User Interface d'une application suit le contexte des conversations dans la durée, sait gérer les interactions optimalement, sait reconnaître ses erreurs, etc.

L'américain **Nuance,** qui fait autour de $2B de chiffre d'affaires, vend sa solution de reconnaissance de la parole (ASR : automated speech recognition) un peu partout en OEM, notamment dans l'automobile. Après avoir intégré des technologies de Nuance dans Siri, Apple a fait l'acquisition de la start-up **VocaliQ** (2011, UK) en 2015. **Sensory** (1994, USA) fait avancer l'état de l'art de manière indépendante depuis plus de 20 ans.

Le couteau suisse **IBM Watson** peut aussi servir à créer sa propre solution pilotée par la parole comme l'a fait l'Américain **Staples** avec son Easy Button qui permet de passer commande de fournitures de bureau.

Citons enfin quelques agents vocaux qui sont destinés aux objets connectés et commercialisés en OEM par leurs créateurs : Snips, LinTO et Mycroft. Ils visent à garantir un meilleur respect de la vie privée des utilisateurs, évitant que les conversations soient envoyées dans le cloud. Une autre solution consisterait à brouiller le signal qui arrive dans les micros des enceintes connectées, mais elle est assez embarrassante[885].

---

[881] Voir Google Home now answers follow-up questions without 'OK, Google' wake word, juin 2018.

[882] Voir This Chatbot Is Trying Hard To Look And Feel Like Us, novembre 2017.

[883] Voir Neon's 'artificial human' avatars could not live up to the CES hype but it wasn't entirely the company's fault par Mat Smith, janvier 2020 et Neon : Samsung lève le voile sur son projet "d'humain artificiel" par Valentin Cimino, janvier 2020.

[884] Voir Google beats Amazon to first place in smart speaker market, juin 2018 qui correspond à la surperfornance de Google sur Q1 2018 et Amazon Takes Top Spot in Smart Speaker Sales in Q2 2018 Says Strategy Analytics, but Google Home Mini Was top Device pour Q2 2018. Puis pour 2019 : Google's Smart Speaker Sales Decline in Q2 2019, Falls Behind Baidu While Device Shipments Rise 55 Percent Globally, Bret Kinsella, août 2019.

[885] Voir This "Bracelet of Silence" actively blocks Alexa and Siri from hearing your conversations par Sarang Sheth, février 2020.

- **Snips** (2013, France, $24,7M) a été acquis par l'américain Sonos fin 2019. Sa stratégie commerciale a donc dû évoluer en conséquence. Il proposait un assistant vocal pour objets connectés qui fonctionne en mode autonome sans nécessiter un aller et retour avec un serveur. Cela servait à mieux respecter la vie privée de l'utilisateur, y compris lorsque le service fait appel à des ressources sur Internet. La solution était commercialisée en OEM et à un prix fixe qui ne dépendait pas du nombre d'utilisateurs. Ils ont développé aussi un SDK matériel intégrant tout ce qu'il faut pour créer son objet connecté à commande vocale. Tout fonctionne sur l'objet connecté qui peut être un kit Raspberry, à base de processeur Intel ou Arm. Il faut un cœur tournant à un minimum de 1GHz, 521 Mo de mémoire. Le système peut être réveillé avec un mot clé (« Wake word »). Les scénarios utilisent de 500 à 2000 mots pour piloter les objets connectés.
- Ils sont déjà partenaires du français **CareOS**, vu au CES 2018, avec son SDK pour équiper la salle de bain en objets connectés et IA, et intégrés dans le robot ménager multi-fonctions d'une autre startup française, **Keecker**, également habituée du CES (mais liquidée en mars 2019). En 2019, CareOS intégrait le support de TensorFlow Lite dans sa plateforme pour le développement de solutions d'IA autonomes pour son miroir connecté Artemis.
- **Mycroft** (2015, USA, $5M) propose une solution embarquée open source. Ils ont aussi leur « reference design » hardware. Mais une partie de leur logiciel nécessite une connexion au cloud. Ils sont partenaires de Mozilla.
- **LinTO** (France) est une filiale de l'ESN Linagora, spécialisée dans l'open source. LinTO est leur agent conversationnel exploitant des ressources en cloud et protégeant la vie privée « by design ». Il se positione comme alternative aux agents conversationnels dominants et marché et tous américains. Ils ont notamment la Société Générale comme client. Ils présentaient au CES 2018 un *reference design* d'enceinte à commande vocale mais cela fonctionne sur d'autres devices comme il se doit. On les retrouvait aux CES suivants, jusqu'au CES 2021 virtuel.

En mai 2019, **Google** réagissait à cette attente d'une plus grande préservation de la vie privée en intégrant une fonction de reconnaissance vocale fonctionnant en local dans Android 10.

Voici quelques autres acteurs de ce marché qui ciblent de plus en plus le marché des centres d'appels aussi bien sortants qu'entrants[886] :

- **Care Angel** (2014, USA, $3,4M) propose un assistant vocal qui joue le rôle d'infirmière virtuelle à destination des patients atteints de maladies chroniques.
- **Mobvoi** (2012, Chine, $252,8M) est une startup de Shanghai proposant un moteur de recherche pilotable par commande vocale. Google Ventures a participé au troisième tour de financement en 2015 avec $60M, lui permettant de mettre un autre pied sur le marché chinois où Google est dominé par Baidu. Le métier principal de cette société est de fabriquer des montres connectées ! Ils ont créé leur Amazon Echo local, le TicHome Mini, de seulement 4,3 cm de haut, 276 grammes et waterproof qui concurrence ceux de Alibaba et Baidu.
- **Smartly.ai** (2012, France, $945K) permet de créer son application vocale facilement. Ils ont déjà la BNP, GRDF et l'Elysée comme clients.
- **Davi** (2000, France) crée des agents vocaux « émotionnels » en collaboration avec le laboratoire LIMSI. Ils sont notamment commercialisés auprès de grandes entreprises françaises.

[886] Voir l'étude La voix monte le son, la révolution des assistants vocaux par Viseo et Roland Berger, 2019 (40 pages).

- **Xbrain.io** (2012, USA) est une startup établie dans la Silicon Valley ainsi qu'à Paris et Lille qui se spécialise dans les applications de l'IA à l'automobile et la robotique. Sa plateforme xBrain Personal Assistant permet de créer des agents conversationnels, utilisés notamment dans les centres d'appels (sous l'appellation satisfaction.ai) et dans l'automobile (vidéo). Elle s'appuie sur la reconnaissance vocale, sur la gestion de contexte, sur la détection des intentions et la gestion de règles. Elle utilise notamment des GAN (generative adversatial networks) pour la génération de dialogues réalistes. Son créateur, Gregory Renard, planche sur l'IA depuis près de 20 ans. Pour lui, le langage est mille fois plus complexe à gérer que la vision. Il faut être créatif pour utiliser l'IA et assembler des techniques très variées : du machine learning à base de SVM jusqu'à des réseaux de neurones à mémoire LSTM.
- **Voysis** (2012, Irlande, $8M) permet de créer des interfaces vocales personnalisées et fonctionnant de manière autonome dans des systèmes embarqués avec ViEW (Voysis Embedded WaveNet) qui exploite le système Wavenet de DeepMind. Cela permet d'ajouter une fonction de recherche vocale dans un site de vente en ligne (vidéo).
- **Dasha AI** (2018, Russie/USA, $2M) se focalise sur les agents conversationnels pour centres d'appels[887]. Ils reproduisent ainsi le scénario de prise de rendez-vous que Google avait présenté en mai 2018 avec leur système Duplex. Mais la vidéo associée, comme pour le cas de Google, fait douter de la réalité artificielle de la voix du service.
- **PolyAI** (2017, UK, $12M) cible aussi les centres d'appels en contact texte et voix.
- **Afiniti** (2006, USA, $197,2M) qui crée des couples entre clients et agents de centres d'appels.
- Et puis un cas d'usage très pointu avec le chatbot **Guru** qui sert à lancer des traitements sur des supercalculateurs[888].

Chatbots textuels

Les robots conversationnels ou "chatbots" sont très en vogue depuis 2015. Des outils permettant d'en créer sont proposés par de nombreuses startups ainsi que dans diverses offres de grandes entreprises du numérique (Facebook, Google, IBM, Facebook, Oracle…).

Les chatbots visent à automatiser le service client en ligne dans les sites de e-commerce, services financiers ou autres. Ils sont utilisés également pour des services internes aux entreprises (gestion de notes de frais, demandes de congés, helpdesk IT ou RH, conseils juridiques…). On les utilise à partir d'applications mobiles, de services de messagerie instantanée comme Facebook Messenger, ou de sites web divers.

L'objectif ultime est de réussir le fameux test de Turing qui définit une intelligence artificielle comme étant une intelligence impossible à distinguer de celle de l'homme dans de telles discussions par le biais d'échanges textuels. On en est encore loin, même avec les chatbots les plus élaborés. Ils sont encore très décevants et pas forcément appréciés des utilisateurs. Ces chatbots peuvent avoir une interface vocale comme avec Google Assistant, Apple Siri, Microsoft Cortana et Amazon Alexa.

Il est assez difficile d'évaluer la maîtrise technologique des différentes sociétés de ce secteur. Elles utilisent un patchwork de différentes APIs et outils de deep learning plus ou moins packagés[889]. Elles peuvent maintenant éventuellement faire appel aux APIs de GPT-3 d'OpenAI qui est entraîné sur un corpus textuel géant.

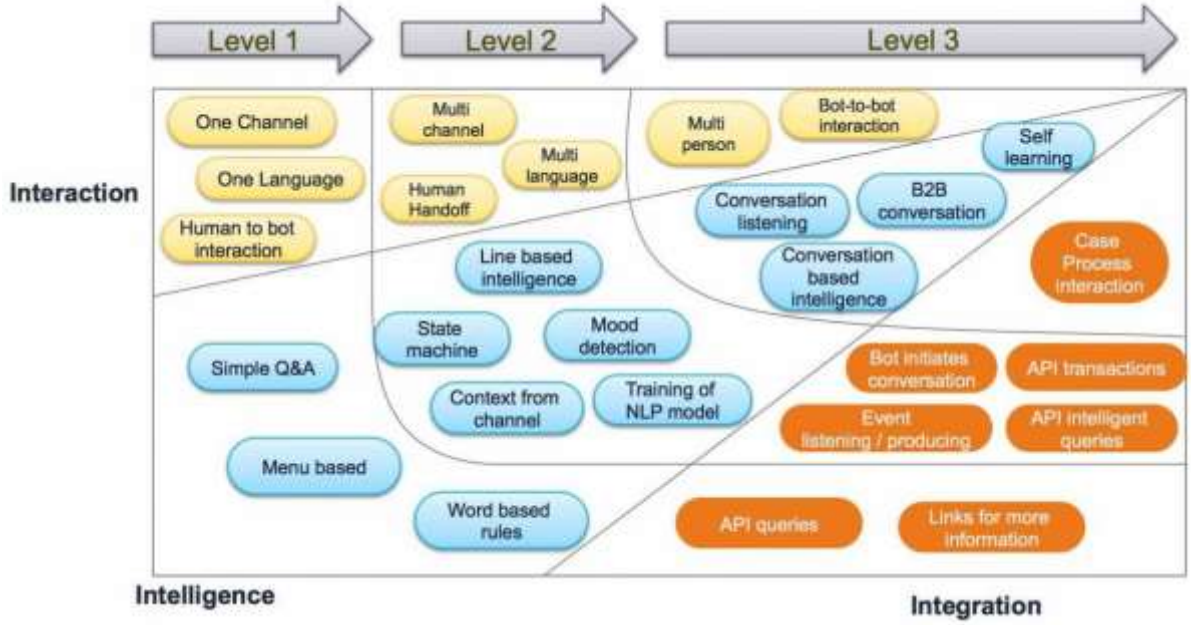
---

[887] Voir Dasha AI is calling so you don't have to par Natasha Lomas, un long article de TechCrunch publié en août 2019.

[888] Voir AI Assistant Runs Supercomputers par Mathew Dirjish en avril 2019.

[889] Cet article très intéressant Contextual Chatbots with Tensorflow de mai 2017 décrit comment développer un chatbot avec le SDK de machine learning et deep learning TensorFlow de Google complété par la bibliothèque TFLearn, le tout étant écrit en langage de programmation Python. Tous ces outils sont open source et gratuits.

Certaines redéveloppent leur propre moteur de traitement du langage, ce qui peut paraître curieux en raison de l'abondance de solutions déjà disponibles sur le marché. D'autres se contentent d'un simple moteur de règles, très rudimentaire dans sa portée[890].



Il existe en effet différentes techniques pour créer un chatbot. Elles vont de la gestion basique de questions/réponses à des bots plus sophistiqués capables de trouver de l'information dans des sources variées, de mener des discussions en mémorisant bien leur contexte et de prendre l'initiative, le tout grâce à des techniques avancées de traitement du langage, à des modèles prédictifs et en tenant aussi compte de l'humeur du client. On appelle cela un chatbot de niveau 3[891].

Dans tous les cas de figure, un bon chatbot doit être alimenté par des sources d'information diverses :

- L'accès à des **applicatifs métiers** divers pour interroger des bases de données, faire des réservations, bref, être intégré dans divers systèmes transactionnels.
- L'exploitation d'outils de communication existants avec les clients comme les logs de centres d'appels, les discussions dans les **réseaux sociaux** d'où l'on pourra extraire des dialogues entre personnes réelles pour identifier des réponses à de nouvelles questions.
- Des **scénarios d'accueil** et de **questions/réponses** (*exemples ci-dessous*) ce qui peut être très laborieux à saisir si cette connaissance n'est pas déjà formalisée dans l'entreprise ou si elle est difficile à capter.

[890] Voir L'arnaque chatbots durera-t-elle encore longtemps ? par Thomas Gouritin, octobre 2017.

[891] Le schéma au-dessus qui décrit les caractéristiques de trois niveaux de chatbots provient de How can Chatbots meet expectations? Introducing the Bot Maturity Model, Léon Smiers, Oracle, avril 2017.

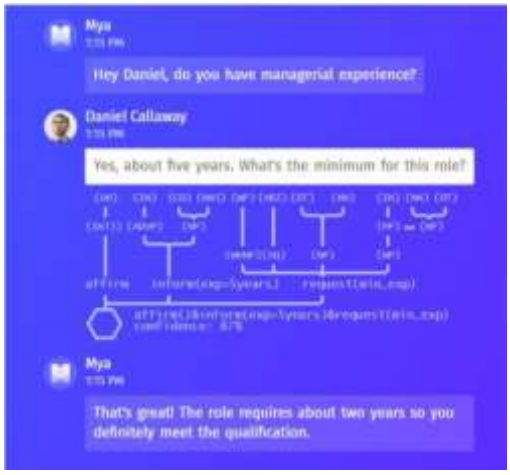
Toutes ces connexions ne se feront pas d'un claquement de doigts ! En général, plus la solution est verticale, moins la startup de chatbot doit disposer de technologie en propre. Ces sociétés se distinguent beaucoup plus par les marchés visés que par leurs choix technologiques ou leurs performances.

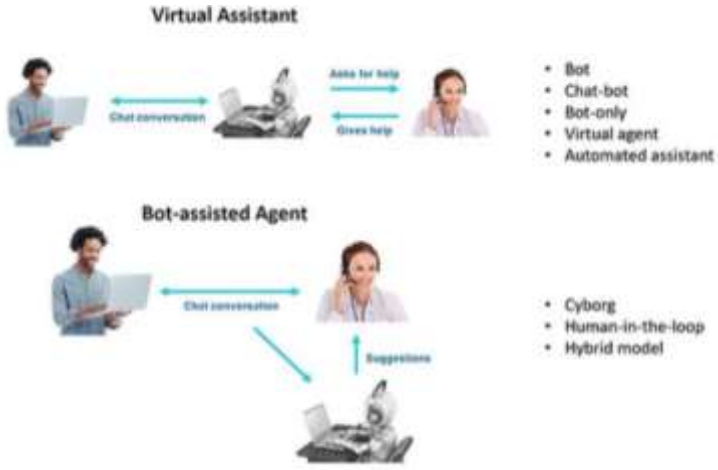
A ce stade de leur développement, les chatbots ne répondent habituellement qu'à des questions très formatées dans un espace sémantique limité au métier de l'entreprise qui le propose. Ils ne savent évidemment pas bien répondre à des questions très ouvertes. Et lorsque la réponse est correcte, il s'agit souvent d'un copier-coller d'une réponse humaine existante dont la grammaire est éventuellement ajustée pour s'adapter au dialogue en cours.

Parfois même, les chatbots génèrent un effet miroir de la bêtise humaine, comme ce fut le cas en 2016 avec le chatbot expérimental de Microsoft Research qui devint rapidement raciste et dû être débranché[892]. En cause, les méthodes d'apprentissage automatiques exploitant des dialogues avec des utilisateurs. Et c'était avant le 8 novembre 2016 ! Heureusement, les chatbots circonscrits à un domaine métier donné risquent moins de se retrouver dans ce cas-là.

Les chatbots sont de trois types différents du côté des interactions avec leurs utilisateurs :

- Ceux qui fonctionnent de manière entièrement **autonome**. Ce sont des assistants virtuels.
- Ceux qui fonctionnent de manière **semi-autonome** et sont animés par des opérateurs humains lorsqu'ils ne savent pas bien répondre.
- Ceux qui **aident** des opérateurs humains à répondre aux questions des clients dans les centres d'appels. Sachant que dans certains cas, il n'y a plus d'humains[893] !





L'offre peut être segmentée avec des chatbots généralistes, des chatbots spécialisés dans des domaines précis (ecommerce, recrutement…) et des outils de création de chatbots[894] et des plateformes d'accueil de chatbots comme Facebook Messenger ou Slack.

Les chatbots sont rarement prêts à l'emploi et nécessitent un travail de personnalisation et de mise en place qui est réalisé par le fournisseur, par un de ses partenaires services ou par l'entreprise cliente elle-même. On voit d'ailleurs émerger des agences de réalisation de chatbots qui s'appuient sur les outils de création de chatbots du marché.

---

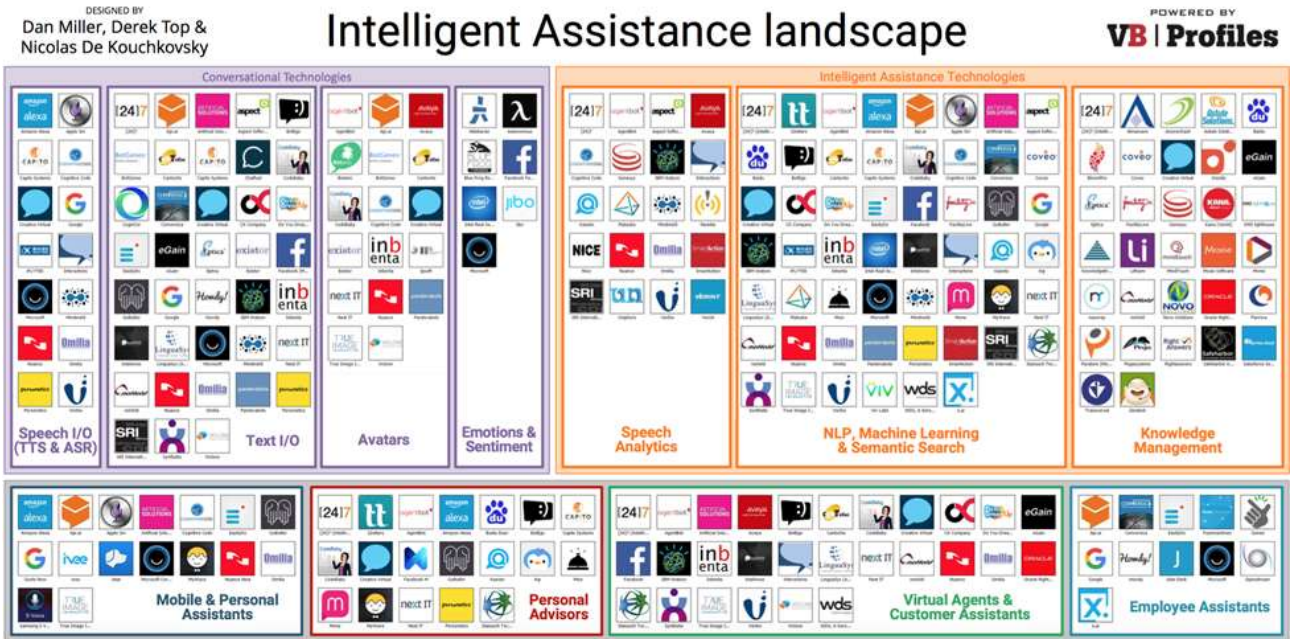
[892] Voir Microsoft is deleting its AI chatbot's incredibly racist tweets, mars 2016.

[893] Voir AirAsia shutters call centres to go all-in on chatbot and voice AI par Asha Barbaschow, décembre 2019.

[894] Voir 25 Chatbot Platforms: A Comparative Table par Olga Davydova, mai 2017, qui recense et compare 25 outils de création chatbots.

Le nombre de startups de chatbots créées ces dernières années est impressionnant. Il rappelle la vague des réseaux sociaux après 2004 et celle des services de vidéo en ligne après l'acquisition de YouTube par Google en 2006 !

Leur diversité témoigne d'un marché en ébullition encore immature. En effet, les marchés matures du numérique se distinguent en général par leur sédimentation autour d'un nombre limité d'acteurs. Cela en prend toutefois la tournure avec quelques leaders qui émergent au niveau des plateformes de chatbots.



La platefome de **Facebook Messenger** domine les usages. En effet, les grandes marques et services l'ont choisie parce que Facebook est le réseau social dominant, en tout cas dans les pays développés. Il est suivi de **Slack**, très utilisé pour le travail collaboratif dans les entreprises.

Nous allons faire ici un panorama de quelques-unes des startups de ce secteur en commençant par quelques plateformes de chatbots généralistes :

- **Semantic Machines** (2015, USA, $20,9M) est une startup de Boston et Berkeley qui propose des chatbots pouvant être intégrés dans toutes sortes d'usages, b2b et b2c. L'équipe fondatrice comprend des anciens de Siri et Google Now. La solution intègre la reconnaissance et la synthèse de la parole.
- **Talla** (2015, USA, $12,3M) propose une solution de chatbots pour les besoins des entreprises, comme dans le recrutement, le marketing et la gestion de rendez-vous. Elle s'intègre dans les systèmes de messagerie tels que Slack. Elle fait penser au français Julie Desk.
- **Dydu** (2009, France), anciennement Do You Dream Up, propose un agent conversationnel multilingues pour les sites web. Il est notamment utilisé par Voyages-SNCF depuis 2011 et a récemment évolué pour être intégré dans une "HelpBox", sorte d'aide en ligne contextuelle interactive. La société a de nombreux clients grands comptes en France, tels qu'EDF dans sa direction juridique, ses hotline IT et RH[895], et plus de cinquante collaborateurs. Elle vient de passer d'une stratégie plutôt services/projets vers une stratégie produit, indispensable pour l'internationalisation de l'activité. En 2019, elle lançait son callbot, un agent conversationnel vocal exploitable dans la relation client. Il est capable de mener une discussion sur plusieurs sujets en même temps et de personnaliser les réponses en fonction du profil client.

[895] Voir EDF créé un chatbot pour transformer la fonction juridique en entreprise, mai 2018 (vidéo).

- **TARA** (2015, USA, $3M) est une startup de San Francisco qui propose un robot conversationnel de gestion du recrutement de freelances.
- **Clustaar** (2013, France, $1,7M) a aussi développé une plateforme en cloud de création de chatbots (vidéo).
- De nombreux chatbots spécialisés avec pour commencer, une palanquée de startups dédiée à la création de chatbots pour les sites de vente en ligne : **Msg.ai** (2014, USA, $7,3M) qui est notamment déployée chez Sony et **MessageYes** (2015, USA, $6,5M), une startup de Seattle, qui associe machine learning et opérateurs humains avec deux spinoffs, l'une qui commercialise des disques vinyles (The Edit) et l'autres, des BD (Origin Bound). The Edit aurait vendu $1M de vinyles en huit mois.

Passons à quelques solutions de chatbots plus originales :

- **IPSoft** (1998, USA) propose son chatbot Amelia qui est positionné sur le helpdesk IT. Elle exploite un agent de détection d'émotion des utilisateurs provenant d'Affectiva. La startup a fait un pivot et propose depuis un chatbot généraliste ciblant la relation client. En 2020, l'éditeur lançait DigitalWorkforce.ai, une place de marché de collaborateurs virtuels, le premier proposé étant, surprise !, un agent de support technique de premier niveau pour les services informatiques (pour le reset de mots de passe et autres dépannages de base). Ils ont un inconvénient notable : des prix très élevés.
- **Existor** (1988, UK) a créé Cleverbot qui exploite la webcam des laptops pour interpréter les visages des utilisateurs. Cleverbot utilise la puissance des GPU des ordinateurs et des mobiles. La société propose aussi un avatar visuel pour mener ces conversations. J'ai fait quelques tests et ce n'est pas très probant (*ci-contre*). Et pour cause, les agents conversationnels sont souvent mis en oeuvre dans des univers sémantiques très précis, comme l'offre d'une société donnée. Ils ne permettent pas de naviguer intelligemment dans Wikipedia par exemple !

- **Davi The Humanizers** (2000, France) fait de même avec son chatbot vocal animé par un avatar.
- **Replika** (2014, Russies/USA, $4,54M) développe son chatbot grand public Replika qui joue le rôle d'un ami, conseiller si ce n'est psychothérapeute. Une sorte de "Her", mais pas encore au point. Il sert surtout à choisir des restaurants. La startup explique que sa solution est basée sur une architecture propriétaire de deep learning et qu'elle est dotée d'un fort quotient émotionnel. La startup a été créée à San Francisco par deux russes, dont un spécialiste du traitement du langage.
- **PocketConfident** (USA) est une startup créée à San Francisco par le Français Olivier Malafronte. Elle développe un agent conversationnel spécialisé dans le coaching personnel. En gros, c'est un psy pas cher. C'est une offre voisine de celle de Replika mais avec un angle un peu différent.
- **Gorgias** (2015, France, $3,4M) est aussi positionnée sur l'automatisation du helpdesk IT. Mais son outil aide les conseillers de support à être plus efficaces, sans les remplacer.
- **Publicis** a développé une application de recommandation de maquillage à base de chatbot pour Sephora[896]. Cela relève encore d'une approche de service sur mesure, pas de la création d'un produit.

[896] Source : keynote de Microsoft AI en septembre 2017, https://myignite.microsoft.com/sessions/56555.

- **Hubware** (2016, France) utilise une approche intrigante en vendant des assistants conversationnels sur mesure sans technologie en propre, en les assemblant selon les besoins du client. Ils apprennent leur métier avec leurs clients, une méthode qui rappelle celle de nombreux cabinets de conseils. A commencer par les sociétés du e-commerce. L'inconvénient de la méthode est que cela rapproche plus la startup d'une société de services que d'une véritable startup à même de générer des économies d'échelle.
- **askR.ai** (2016, France) a développé un chatbot de Business Intelligence servant à interroger des bases de données en langage naturel (vidéo). Reste à voir si cela fonctionne de manière générique pour toutes les bases de données métier (il s'interface avec SAP) et si c'est plus rapide que la manipulation de données dans un outil de business intelligence traditionnel.

Puis aux chatbots associant automatisation et intervention humaine :

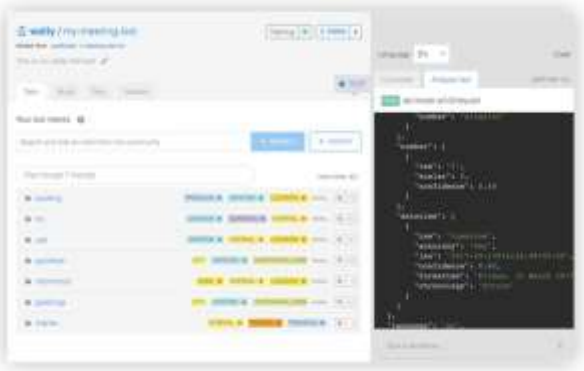
- **Curious.ai** (2013, USA, $7,35M) commercialise DigitalGenius qui associe deep learning et intervention humaine pour les chatbots de services clients. Le chatbot qui fonctionne en mode texte sur site web, réseaux sociaux et SMS est entraîné avec des transcriptions d'appels réels au service client.

- **Julie Desk** (2014, France, $3,3M) propose un agent qui répond de son côté aux mails pour l'organisation de rendez-vous. Lui aussi est supervisé par de vraies personnes pour le contrôle qualité. Julie Desk a un concurrent américain, **x.ai** (2014, USA, $44,3M). En 2019, la startup lançait Slash, une version entièrement autonome de son assistant de gestion de rendez-vous (vidéo).
- **Kwalys** (2014, France) est une autre plateforme de création de chatbots. Elle mélange arbre de décision et deep learning pour trouver les réponses. Ils sont partenaires d'IBM Watson.
- **Botmind** (2017, France) associe aussi chatbot et humains pour répondre aux clients. Leur solution est notamment déployée chez Axa Banque, BNP Paribas, La Redoute et Prixtel (d'après leur site).
- **iAdvize** (2010, France, $55M) propose son agent conversationnel ibbü. Celui-ci est doté depuis 2020 d'une « intelligence augmentée » qui analyse l'ensemble des données clients des marques et le contenu de leurs conversations en ligne. Cela permet d'en extraire les intentions. Le système est aussi amélioré de manière continue par des humains. Il permet de détecter les visiteurs des sites en déterminant ceux qui ont le plus besoin d'une assistance via messagerie instantanée. Le système transfère automatiquement les demandes aux agents appropriés (bot ou conseiller humain). Il peut aussi servir d'outil d'assistance des agents des centres d'appels en leur fournissant des suggestions de réponses et de traduction instantanée.

Et enfin, voici quelques outils de création de chatbots exploitables par les entreprises et agences spécialisées en création de chatbots :

- **Opla.ai** (2015, France), propose aussi un outil de création de chatbots. Une partie des logiciels de traitement du langage a été créée par son cofondateur, Mik Bry.

- Le BotMaker de **Viseo** (1999, France, $2M) est un outil de création de chatbots textes et vocaux en open source doté d'une interface graphique interactive capable notamment d'interroger les systèmes d'information des entreprises. Il s'adapte aux grandes plateformes de chatbots comme Facebook Messenger, Slack, Wechat, Amazon Echo et Cortana. Viseo est une société de services spécialisée initialement dans le déploiement d'ERP et qui est devenue un généraliste de la transformation digitale, avec 1200 collaborateurs et 130M€ de CA.
- **Chatfuel** (2016, USA, $120K) est une jeune startup américaine qui permet de créer ses propres chatbots. Sa solution serait déployée chez Forbes, Techcrunch et dans la messagerie instantanée Telegram qui compte plus de 100 millions d'utilisateurs.
- **Pandorabots** (2008, USA) est une startup d'Oakland (Californie) qui propose une plateforme de chatbots en ligne, open source et multi-lingue. 300 000 chatbots avaient été générés au décompte de septembre 2017. Ils sont intégrables dans divers environnements de messagerie instantanés tels que Slack et Whatsapp.
- **Recast.ai** (2015, France, $2,25M) est une startup française créée par des anciens de l'école 42 qui propose un outil de création de chatbots et un SDK associé. L'ensemble est très bien packagé et s'utilise en mode cloud (*ci-dessous*). Au cœur de leur solution se trouvent différentes briques internes et externes. L'équipe a pris soin de développer certaines briques de traitement du langage en interne. La startup a été acquise par SAP en janvier 2018 et sa solution sera intégrée dans SAP Leonardo Machine Learning.



- **Viv Labs** (2012, USA, $30M) est une startup californienne qui propose les outils de création d'assistants vocaux avec des fonctionnalités voisines de celles de SIRI.
- **Expert System** (1989, Italie, $32,7M) est une société spécialisée dans le traitement du langage qui exploite une approche hybride associant analyse sémantique, graphes et machine learning. On pourrait simplifier en disant qu'il s'agit d'un mélange d'IA symbolique et connexionniste. Elle propose des briques logicielles de sa famille Cogito pour gérer des contenus non structurés qui automatisent le traitement d'informations dans les services clients, pour la conformité, l'atténuation des risques, les applications de renseignement et de veille. Leur approche permettrait de créer des solutions d'IA explicables. Leur offre produit cible en particulier le marché des assurances (Cogito for Claims, Cogito for Underwriting) et celui de la santé (Clinical Research Navigator, Medical Intelligence Platform, Cogito for Life Science)[897].
- Et bien évidemment, les solutions du domaine issues des GAFAMI, notamment Messenger chez **Facebook** qui permet de développer son propre chatbot, les outils développeurs de Cortana chez **Microsoft,** d'**IBM Watson**, de **Google**[898] et **Oracle**.

---

[897] Voir leur livre blanc : Cogito, a unique artificial intelligence technology, 2018 (13 pages).

[898] DialogFlow est l'outil de création de chatbots de Google. Voir Les concepts de base de Dialogflow pour écrire votre Chatbot par Xavier Ardisson, novembre 2018 et Introduction aux Chatbots avec Dialogflow par Hao Chen, mai 2018.

- **Viv** (2012, USA, $30M) acquis par Samsung en octobre 2016, propose un outil de création de chatbot.

En voici d'autres encore :

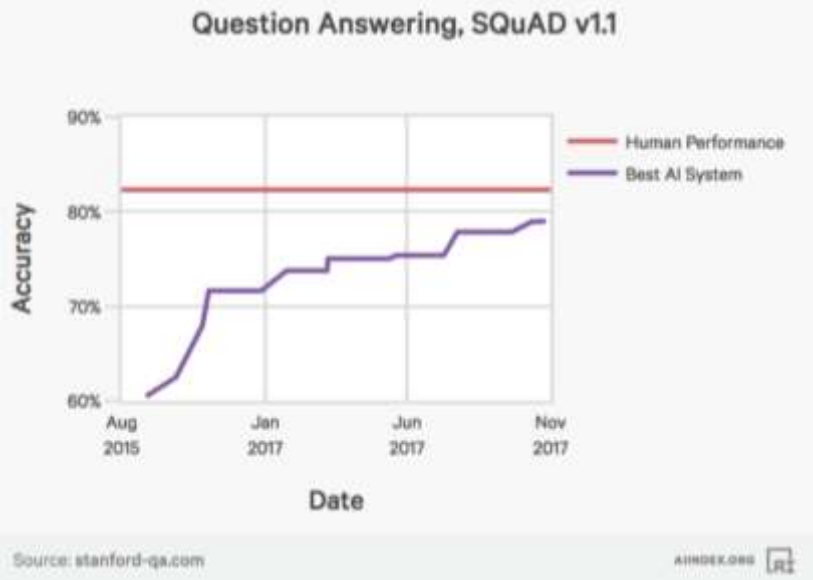
- **Botfuel** (2016, France, $4M) propose un SDK de création de chatbots pour développeurs en entreprise.
- **Botnation** (2016, France), basée à Nantes, propose une plateforme en ligne de création de chatbots pour les TPE/PME. Elle a déjà 3500 entreprises utilisatrices. La spécificité ? Des fonctions marketing de tracking.
- **Braina** (2013, Inde) offre la panoplie complète de création de chatbot textuel ou vocal. Cette startup m'a donné du fil à retordre car sur son site et dans la Crunchbase, impossible de savoir d'où elle venait. C'est finalement dans LinkedIn que j'ai découvert que la société était basée en Inde. Eux aussi ont peur de leur ombre !
- **Golem.ai** (2016, France, 1,3M€) a développé une technologie de création de chatbots originale qui nécessite moins d'entraînement que les solutions classiques et qui peut fonctionner de manière autonome. **Ils** utilisent un mix d'IA connexionniste et d'IA symbolique et la solution générée serait explicable.
- **Omni Ai** (2017, USA, $45M) est un autre original du chatbot qui a développé une IA à base de deep learning autosupervisée, on se demande comment. Et en même temps, ils font aussi de l'analyse de vidéos, suite à l'acquisition de la startup BRS Labs[899].
- **Konverso** (2017, France) a une offre de création de chatbots pour la fonction support des entreprises.
- **LogMeIn** (2003, USA, $30M) propose le chatbot de relation client Bold360 au sein d'une large gamme de logiciels de travail collaboratif pour les entreprises et services publics. Son chatbot s'intègre dans les applications de messagerie instantanées courantes du marché (WhatsApp, Facebook Messenger, WeChat et SMS) pour servir directement les clients ou les collaborateurs de l'entreprise en contact avec ces derniers. Le système supporte l'escalade de demandes vers des opérateurs humains. Il gère aussi la répartition des flux entrants de mails clients.
- **Wiidii** (2014, France) propose un assistant hybride, avec IA maison et assistants personnels humains. Ils ont signé avec un constructeur allemand haut de gamme et ont une application mobile b2c permettant de tester ses capacités.
- **Arvato** propose aussi une offre de service intégrée pour création d'agents conversationnels, en liaison avec des centres d'appels.

Il existe même des prix récompensant les chatbots s'approchant le mieux du test de Turing ou le passant entièrement : les **Loebner Prizes**, créés en 1990. S'il a bien été attribué chaque année depuis dans sa première mouture, et notamment au créateur de Cleverbot en 2005 et 2006, il ne l'a pas encore été dans la seconde, celle du passage complet du test du Turing devant deux juges[900].

[899] Voir la présentation OMNI AI (27 slides).

[900] En anglais, des tests de compréhension du langage sont aussi réalisables avec l'outil de benchmarking SuperGLUE. Voir SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems, NUY et Facebook, 2019 (30 pages).

Les créateurs de chatbots généralistes peuvent entraîner leur créature avec **Base Squad 2.0** qui est une base de questions/réponses créée en mode crowdsourcing par l'Université de Stanford [901] . SquAD2.0 combine 100K questions qui ont des réponses et 50 000 questions qui n'ont pas de réponse. Elles doivent permettre de réaliser des tests de chatbots pour vérifier qu'ils ne répondent pas n'importe quoi aux questions qui n'ont pas de réponse et répondent bien aux autres questions.



SquAD a été utilisé par des chercheurs de Microsoft dans **ReasoNet**, une technique associant IA symbolique et connexionniste[902]. Début 2018, Alibaba et Microsoft faisaient réussir le test SquAD par leurs chatbots respectifs, et battre de très peu les capacités humaines[903]. C'est illustré dans le graphe *ci-dessus* qui montre la vitesse de progression dans la création de solutions de réponses au test SQuAD (ici en version 1.1)[904]. Alibaba améliorait encore son système en 2019 avec **Alime**, leur chatbot de service client servant 50 millions d'utilisateurs par jour pour ses sites Taobao et Tmall. Il était capable de traiter 98% des demandes clients selon Alibaba[905].

Dans les chatbots originaux, il y a le cas d'un chatbot développé pour l'église d'Angleterre à la suite d'un concours organisé par cette dernière en 2017[906] !

Voici une synthèse de méthodologie pour réaliser son chatbot[907] :

- **Clarifier les objectifs** : s'agit-il d'améliorer la satisfaction client, de rendre le service client disponible hors des heures d'ouverture du centre d'appel, est-ce pour traiter automatiquement une partie des flux entrants voix+email, pour cartographier les besoins des utilisateurs, pour recruter de nouveaux clients, pour faire du support technique ?
- **Définir un processus** : quelles sont les interventions humaines avant ou après le chatbot, quels sont les scénarios de dialogue type, les processus d'escalade, les tests, les indicateurs de qualité ?
- **Cahier des charges** : définir le niveau d'interactivité et de couverture métier du chatbot ainsi que les interactions possibles avec les applicatifs métiers de l'entreprise en mode transactionnel.

---

[901] Voir Know What You Don't Know: Unanswerable Questions for SQuAD, juin 2018 (9 pages).

[902] Voir ReasoNet: Learning to Stop Reading in Machine Comprehension, 2016 (9 pages) et décrite dans Machine Reading for Question Answering: from symbolic to neural computation de Jianfeng Gao, Rangan Majumder et Bill Dolan, juillet 2018 (31 slides).

[903] Voir Microsoft, Alibaba AI programs beat humans in a Stanford reading test, janvier 2018. Il semble que les chatbots qui ont réussi cette prouesse aient été réalisés sur mesure et ne sont pas les chatbots commerciaux d'Alibaba et Microsoft. On est encore dans un scénario d'AI étroite !

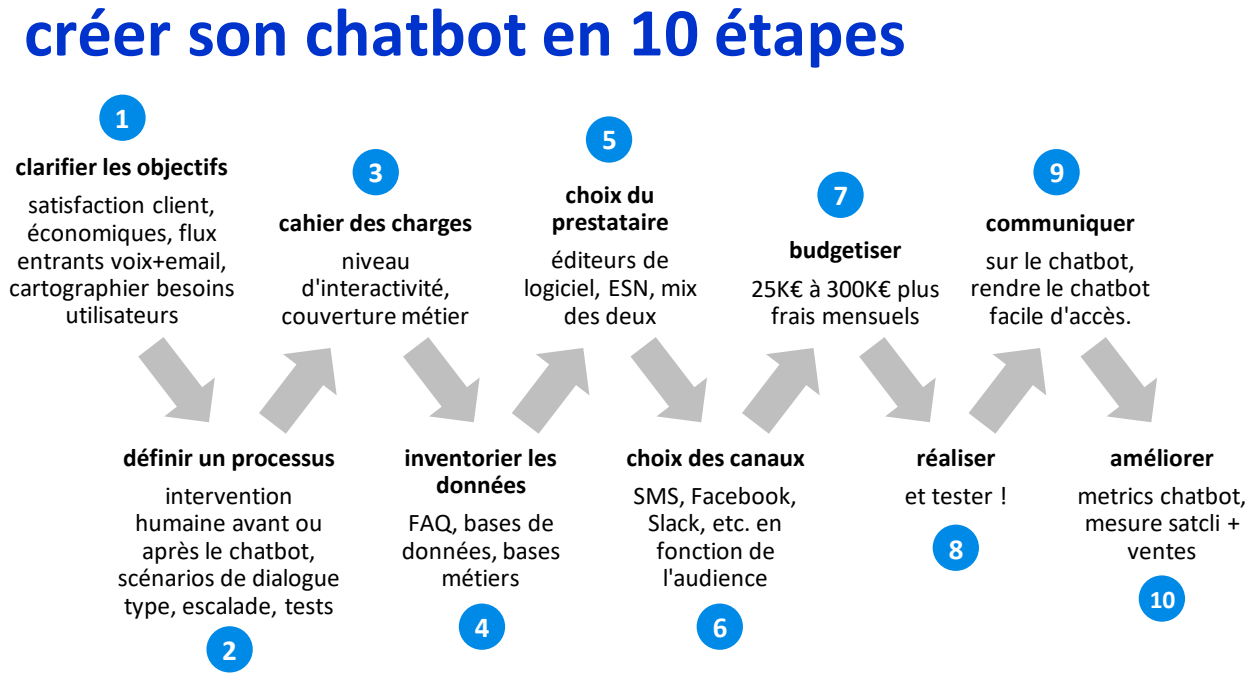
[904] Issu de Artificial Index Report 2017 (101 pages).

[905] Voir Alibaba AI Beats Humans in Reading-Comprehension Test par Christine Chou, juillet 2019.

[906] Voir Chatbot for churchgoers among winners of first Church of England Digital Labs competition, octobre 2017.

[907] Je me suis inspiré de plusieurs sources d'information dont Comment réussir votre projet Chatbot - Twelve Consulting par Glenn Cheynier, 2017, 10 étapes clés pour réussir votre projet Chatbot, juin 2018 et Comment réussir son chatbot ?, 2018.

- **Données** : quelles sont les données qui pourront alimenter le chatbot ? Comment collecter les FAQ, les bases de données, les bases de données métiers, les interactions humaines dans les réseaux sociaux, … Comme aussi les mettre à jour à la volée lors de l'exploitation.
- **Choix du prestataire** : faut-il partir avec un éditeur de logiciel, une startup du domaine, une entreprise de service numérique, un mix des trois ?
- **Choix des canaux** : via quelques canaux numériques le chatbot sera-t-il disponible ? SMS, Facebook, Slack, etc. en fonction de l'audience.
- **Budget** : prévoir 25K€ à 300K€ de frais de mise en route plus des frais mensuels puis des coûts de maintenance dans la durée.
- **Communiquer** : comment communiquer sur le chatbot auprès de sa base d'utilisateurs et le rendre facile d'accès.
- **Amélioration continue** : définir les indicateurs de succès, vérifier l'impact sur la satisfaction client, les ventes et les réachats.



Traduction automatique

La traduction automatique s'est longtemps appuyée sur des méthodes statistiques avec énormément de bidouillage manuel. Mise en œuvre à partir des années 1990, elle s'appuyait sur des bases de données, avec d'un côté une base de traduction contenant des expressions et phrases dans deux langues avec des probabilités de correspondance et de l'autre, une base linguistique contenant des textes dans la langue cible corrects d'un point de vue grammatical et du style.

Le deep learning a fait son apparition dans le domaine relativement récemment, vers 2012. Il exploite des réseaux de neurones récurrents (RNN), leur variante à mémoire (LSTM : Long Short Term Memory) et de nombreuses autres déclinaisons[908].

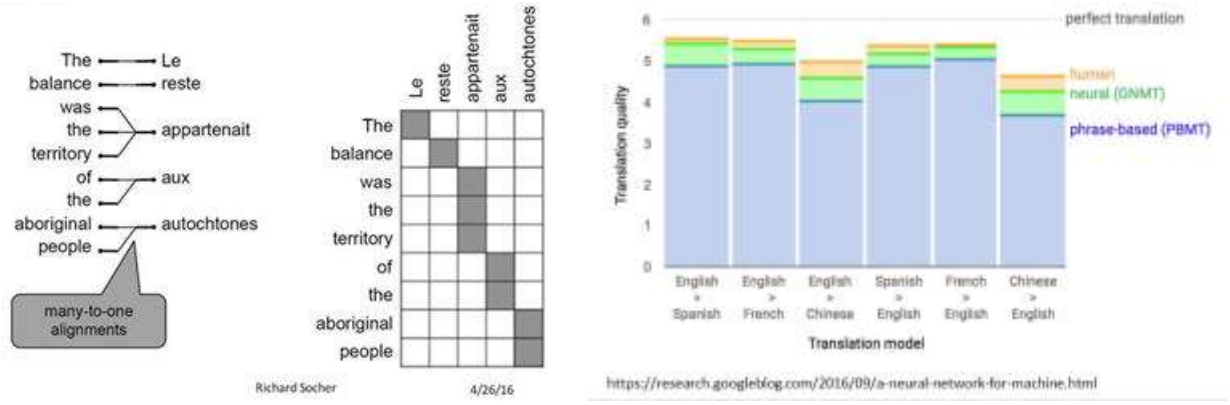
Ces nouvelles techniques permettent de mieux gérer la traduction en respectant le contexte des mots, expressions et phrases. Cela a permis de passer du presque mot à mot à phrase à phrase, en tenant compte du contexte.

[908] Voir la conférence Traduction et traitement de la langue naturelle d'Huggo Schwenk dans la chaire de Yann Le Cun au Collège de France en avril 2016.

Ce champ d'application s'appelle en américain le Neural MT pour **Neural Machine Translation**. Ce courant est devenu dominant en 2016. Il reste basé sur des méthodes probabilistes et s'améliore régulièrement[909]. Sa qualité dépend notamment des bases d'entraînement.

Contrairement à la reconnaissance d'images où l'IA a dépassé les capacités humaines, la traduction à base d'IA n'y est pas encore, tout comme les approximatifs chatbots. La traduction à base d'IA est encore imparfaite mais elle réalise des progrès constants, les langues asiatiques étant toujours plus difficiles à gérer car elles sont plus imagées que les langues européennes. D'où la performance remarquée de Rick Rashid, à l'époque patron de Microsoft Research, lorsqu'il démontra en Chine une solution de traduction orale de l'anglais au chinois en 2012[910].

Les systèmes de traduction les plus sophistiqués sont ceux qui font du speech-to-speech, à savoir qu'ils interprètent la voix et non du texte et le transforment en voix dans la langue cible. Ils alignent donc au minimum trois agents : speech-to-text, traduction puis text-to-speech.



Ce dernier agent peut d'ailleurs lui aussi s'appuyer sur du deep learning pour générer une voix aussi réaliste que possible[911].

L'un des leaders mondiaux indépendants de la traduction est **Systran**. Cette société américaine créée en 1968 avait démarré en traduisant du russe en anglais pendant la guerre froide. Elle est devenue française en 1986 puis acquise par le coréen **CSLi** en 2014. Elle faisait moins de 10M€ de chiffre d'affaires en 2009. En 2018, l'éditeur lançait Pure Neural Server, une nouvelle génération de solution de traduction automatique destinée aux entreprises construite autour du framework open source maison de deep learning **OpenNMT** lancé en 2016. Ils ont aussi une offre spécialisée dans la traduction de documents financiers pour la mise en conformité avec la réglementation européenne MiFID II (Markets in Financial Instruments Directive II).

Google et Microsoft proposent chacun de leur côté un système de traduction automatique avec l'application mobile Google Translate d'un côté et Cortana de l'autre.

Google Translate a la particularité de traduire le texte photographié dans des images.

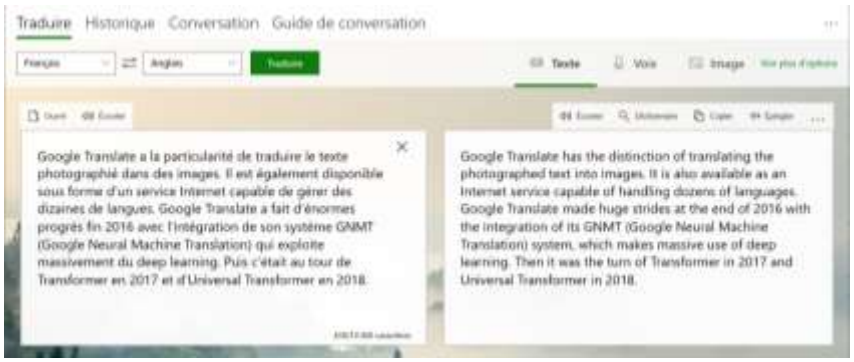
Il est également disponible sous forme d'un service Internet capable de gérer des dizaines de langues.

[909] Voir La traduction automatique fait des pas de géant par Stefano Lupieri, Les Echos, mars 2019.

[910] Visualisable ici : https://www.youtube.com/watch?v=Nu-nlQqFCKg.

[911] La méthode est documentée dans la présentation Deep Learning in Speech Synthesis de Heiga Zen de Google, 2013 (48 slides).

Google Translate a fait d'énormes progrès fin 2016 avec l'intégration de son système GNMT (Google Neural Machine Translation) qui exploite massivement du deep learning. Puis c'était au tour de Transformer en 2017 et d'Universal Transformer en 2018.



Leur particularité est d'accélérer l'entraînement du réseau de neurones en parallélisant les tâches pour chaque mot dans leur contexte, et d'améliorer le résultat de la traduction. Il passe plus de temps à traiter le cas des mots ambigus dont le sens dépend le plus du contexte de la phrase[912].

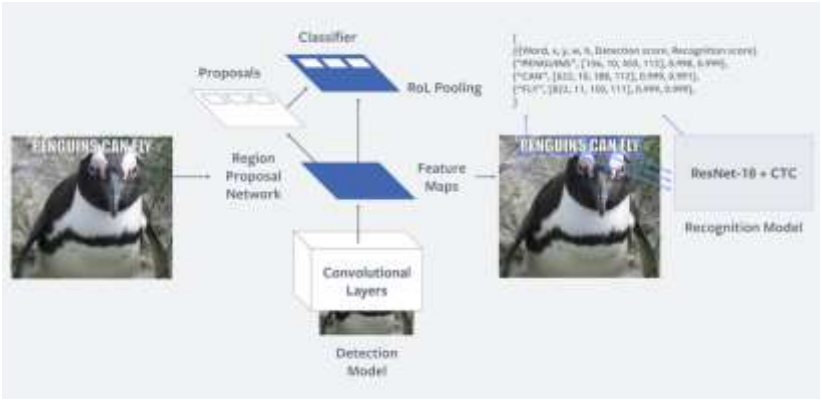
Microsoft Cortana est notamment disponible sous la forme d'une application Windows 10. Sa traduction utilise plus d'idiomatismes anglais comme l'exemple *ci-dessus* en bas peut l'illustrer.

Cortana propose aussi du speech-to-speech. Microsoft teste depuis le printemps 2018 une version de son système de traduction fonctionnant en local sur smartphones, notamment sur les Huawei et leur processeur Kirin 970[913].

Translatotron est un système de traduction speech-to-speech de Google qui évite de passer par un format intermédiaire textuel. Il permet aussi de transférer le timbre de la voix d'une personne dans sa voix traduite dans une autre langue[914]. C'est pour l'instant un prototype testé entre l'espagnol et l'anglais.

Facebook a développé Fairseq, un SDK open source de traduction[915] qui s'appuie sur un réseau convolutif au lieu d'employer des réseaux récurrents comme les systèmes de deep learning habituels (même si nombre de laboratoires et startups s'en éloignent de plus en plus).

Les équipes de recherche du FAIR [916] ont notamment créé Rosetta, une technique qui améliore la qualité de la traduction en exploitant des images contenant du texte[917]. C'est une forme de deep learning multimodal comme celui qui est utilisé pour la reconnaissance de la parole exploitant la voix et la vidéo du locuteur que nous avons vu chez Google.



[912] Les détails sont dans Moving Beyond Translation with the Universal Transformer, août 2018, mais je ne vous cache pas que je n'ai pas tout compris !

[913] Voir Microsoft Translator gets offline AI translations de Frederic Lardinois, avril 2018.

[914] Voir Google's AI can now translate your speech while keeping your voice par Karen Hao, mai 2019 et Voir Introducing Translatotron: An End-to-End Speech-to-Speech Translation Model par Ye Jia et Ron Weiss, mai 2019.

[915] Voir Announcing Fairseq, mai 2017.

[916] Facebook AI Research, dont la direction scientifique est assurée par le Français Yann Le Cun.

[917] Voir Rosetta: Understanding text in images and videos with machine learning, septembre 2018.

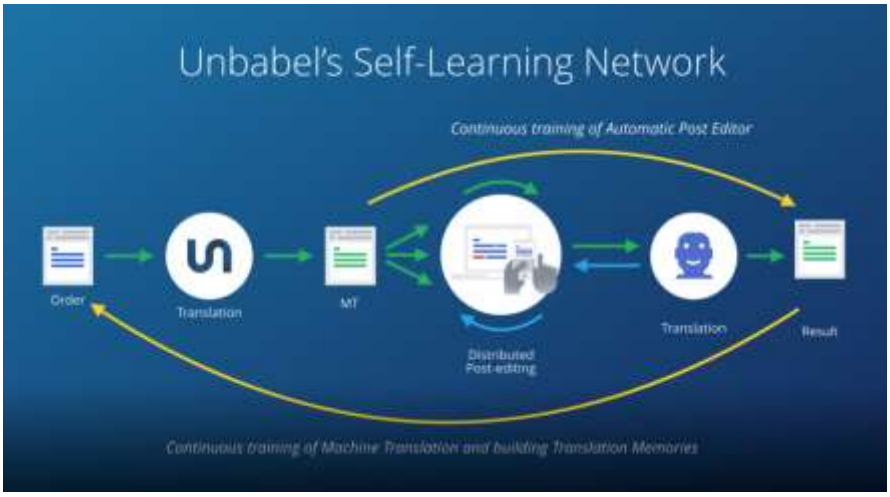
Baidu est aussi un acteur de ce créneau avec des avancées constantes de sa recherche dans la traduction simultanée. Récemment, leurs laboratoires ont démontré un système de traduction qui réduit le temps de latence de la traduction grâce à un modèle prédictif des mots qui vont être prononcés par le locuteur à traduire. Avec leur STACL, la traduction n'a plus qu'une latence de cinq mots avec un taux d'erreur acceptable. Le système gagne en fait environ 2 à 3 mots de latence ce qui n'est pas grand-chose mais peut servir dans la traduction simultanée dans les instances internationales et dans des contextes politiques sensibles[918]. C'est un projet de recherche, donc pas encore opérationnel.

Quittons les grands acteurs du numérique et passons à quelques startups du domaine de la traduction. Nombreuses sont celles qui s'appuient sur du crowdsourcing pour améliorer leurs bases ou pour fournir un service de traduction humain assisté par des outils[919].

Lilt (2015, USA, $3M) propose un système de traduction destiné aux traducteurs professionnels.

L'outil de saisie prédictif propose des suggestions au traducteur pendant son travail. Comme le traducteur ajuste le texte qui est proposé, cela permet en retour d'alimenter le système, ce qui constitue une forme d'apprentissage par renforcement.

Unbabel (2013, Portugal, $31M) associe le deep learning et l'intervention humaine, qui améliore de manière continue le système. La startup indique utiliser une communauté de 50 000 traducteurs indépendants couvrant 28 langues. Leur outil s'intègre dans Salesforce et Zendesk et dans d'autres logiciels métiers via leur API.



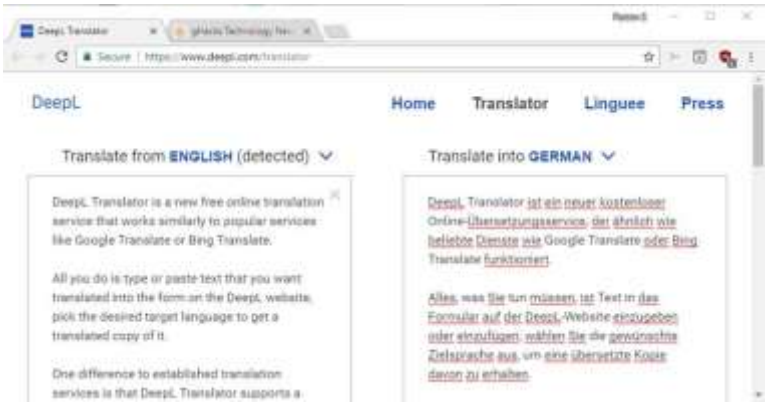
[918] Voir STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency, octobre 2018 (10 pages).

[919] Ces startups sont extraites de 7 Machine Language Translator Startups, mars 2018.

L'intervention humaine peut être vue à la fois comme un moyen d'améliorer la qualité de la prestation et une limitation intrinsèque de leur IA. On est en droit de douter.

SDL (1992, UK) est un acteur établi de la traduction avec un CA de $490M qui propose, entre autres choses, une solution serveur de traduction pour les entreprises, à base de deep learning. La société fournit aussi des services de traduction traditionnels, ou assistés par ses propres outils.

DeepL (2009, Allemagne) s'appuie sur un supercalculateur pour traduire un million de mots par seconde dans 7 langues avec des résultats qui dépasseraient ceux de Google en qualité. Au printemps 2018, DeepL s'associait avec **Quantmetry** (2010, France) pour traduire un ouvrage scientifique de 800 pages, de l'anglais vers le français, en deux mois et demi, corrections humaines comprises[920].



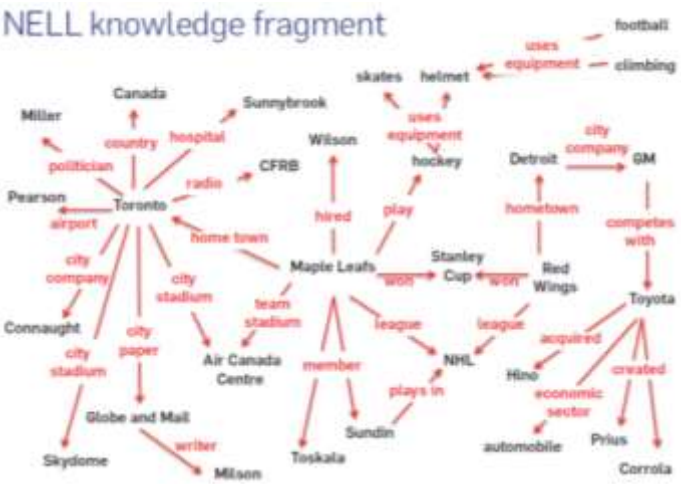
Côté études de cas, signalons le cas de **BNP Paribas** qui a lancé Translate, un outil de traduction automatique destiné à de la documentation métier de toute sorte. L'outil a été créé par une équipe interne de la banque, constituée de data scientists, de développeurs en machine learning et généralistes. Il gère 50 000 traductions par mois dans une vingtaine d'activités. Le tout est hébergé dans le cloud privé interne de la banque sur des serveurs dotés de GPU Nvidia.

Extraction de données

Le traitement du langage a d'autres applications diverses consistant à exploiter les données qualitatives ou quantitatives extraites de textes.

Ces applications permettent d'identifier des personnes, des sociétés, produits, lieux, prix ou dates dans des textes et notamment sur Internet et dans les réseaux sociaux. Elles classifient ces informations selon divers critères comme les sujets de discussion ou la tonalité, que l'on appelle l'analyse de sentiments. Les techniques d'IA sont de plus en plus courantes dans ces applications.

Dans ce domaine, l'une des initiatives les plus marquantes est le projet **NELL** (Never Ending Language Learning) est un service d'identification des liens sémantiques entre expressions textuelles lancé par Carnegie Mellon en 2010. Il scanne automatiquement des millions de sources de textes, notamment sur Internet, pour créer sa base. Le projet a été cofinancé par la DARPA et Google. Il génère une sorte de base de contenus sémantiques reliés entre eux. Comme son nom l'indique, le projet n'est jamais terminé et continue sans cesse de s'enrichir.



[920] Il s'agissait de l'ouvrage de référence « Deep Learning" de Ian Goodfellow, Yoshua Bengio et Aaron Courville. Mais deux mois et demi, c'est encore trop pour traduire dans le sens inverse le Rapport du CES en anglais !

Les datasets (en anglais) sont librement téléchargeables pour être exploités dans des applications de traitement du langage[921]. Ils comprennent plus de 2,4 millions d'entités.

L'exploitation de réseaux de neurones à mémoire (RNN et LSTM) et leurs variantes et combinaisons permet aussi d'améliorer la capacité à détecter les sentiments dans des textes, de manière plus fine qu'avec une simple analyse syntaxique classique[922]. Ces techniques servent notamment à détecter les discours de haine sur Internet[923].

De nombreuses startups et entreprises opèrent dans ce secteur.

IBM propose de nombreux outils de traitement du langage dans sa boîte à outils Watson, dont DeepQA, qui permet non seulement de créer des agents conversationnels, mais aussi d'extraire des règles issues de textes, comme des documents scientifiques. Cela permet d'alimenter en retour des moteurs de règles pour du raisonnement automatique.

Cinnamon (2012, Japon, $10M) fait aussi de l'extraction d'informations de documents qui permet à leur "Lapis Engine" de faire des recommandations ciblées. Ils ont aussi un "Flax Scanner" qui scanne des documents et extrait les informations non structurées qu'ils contiennent.

Heuritech (2013, France, 5,2M€) propose sa solution logicielle Hakken d'analyse sémantique, de tagging et classement automatiques de textes, images et vidéos sous forme d'APIs.

Ils proposent aussi HeuritechDIP qui permet d'améliorer sa connaissance des clients et d'anticiper leurs besoins, évidemment, surtout dans les applications de commerce en ligne et le fashion. Le tout exploite force marchine et deep learning. La startup s'appuie sur les travaux de recherche du CNRS LIP6 et de l'ISIR de l'UPMC (Paris VI). Les deux fondateurs ont un doctorat en IA. Ce n'est donc pas de l'IA washing a priori !

Proxem (2007, France, 1M€) propose Proxem Studio, une solution de traitement automatique du langage permettant de filtrer, analyser, tagger et classifier automatiquement de gros volumes de données textuelles, comme dans les commentaires d'utilisateurs dans les réseaux sociaux ou sites de e-commerce. Le tout s'appuie sur des techniques de machine learning et de deep learning. L'outil permet notamment d'explorer les données analysées de manière visuelle pour identifier des patterns et signaux faibles. Elle s'est fait remarquer en 2016 en étant utilisée par l'équipe de campagne d'Emmanuel Macron pour analyser le contenu des enquêtes terrain et faire ressortir les thématiques clés.

Keluro (2014, France), créée par des anciens de l'ENS, se focalise de son côté sur l'exploitation des emails d'entreprises pour en tirer des informations exploitables et structurer les conversations. Ils exploitent des techniques de machine learning pour la classification des informations. La solution est disponible dans Google Play et sous forme d'extension de Microsoft Office.

Qemotion (2015, France, 614K€) extrait et analyse les réactions textuelles d'une audience pour les convertir en analyses des sentiments.

---

[921] Voir Aiming to Learn as We Do, a Machine Teaches Itself de Steve Lhor, 2010, Never-Ending Learning, mai 2018.

[922] Voir OpenAI sets benchmark for sentiment analysis using an efficient mLSTM, avril 2017. On peut aussi combiner des réseaux convolutifs (CNN) et à mémoire (LSTM) puis faire la moyenne des résultats comme dans Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models par Shervin Minaee & Al, avril 2019 (6 pages).

[923] Voir Internet researchers harness the power of algorithms to find hate speech, octobre 2017.

Klaxoon (2014, France, $55,6M) analyse les textes des idées produits par les groupes de travail dans leur outil de travail collaboratif Brainstorm, afin de proposer des actions pertinentes associées, comme la création de rendez-vous automatique. Klaxoon ne s'intègre pas au passage avec Microsoft Teams, la solution de travail collaboratif de l'éditeur américain. L'annonce en était faite en session plénière lors de l'événement Microsoft Experience à Paris le 7 novembre 2018.

Lexistems (France) propose une API, X-Act.ai, qui permet l'interrogation et le traitement de données et documents par le sens. Elle permet de consolider facilement des informations et analyses à partir de sources de connaissances multiples et hétérogènes. Je les ai croisés à Nantes au Web2day 2018 ! Cela serait sympathique que leur solution puisse permettre de produire automatiquement le Rapport du CES ou ce genre d'ebook !

Cortical.ai (2011, Autriche, $6,3M) propose un outil permettant d'exploiter ses bases textuelles pour y extraire des données, faire des recherches et des annotations.

re:infer (2015, UK, $3,5M) analyse tous les flux textuels et de communication de l'entreprise pour identifier les actions à gérer dans la relation client. Ils sont notamment fournisseur de la solution de RPA **Blue Prism**.

Tessian (2013, UK, $16,8M) est positionnée sur un créneau voisin de celui de Keluro. Ils analysent les emails sortant des entreprises pour bloquer l'envoi d'informations qui pourraient être sensibles (confidentielles, mauvais destinataire…).

SkipFlag (2014, USA) exploite avec Rover les données circulant dans les entreprises notamment dans les outils collaboratifs comme Slack pour créer des bases de connaissances capables de répondre aux questions clés. Encore une application exploitant l'analyse du langage.

Wrapidity (2015, UK) a développé une technologie pour automatiser l'extraction de données à partir de contenus Web non-structurés. Elle a été acquise par la société de datal analytics **Meltwater** (2014, USA, $60M) début 2017.

Idibon (2012, USA, $6,9M) analyse les textes structurés, notamment issus des réseaux sociaux, pour les classifier automatiquement et réaliser des analyses statistiques dessus.

WorkFusion (2010, USA, $121M) propose l'automatisation de l'exploitation de gros volumes de données non structurées. Il donne l'impression de récupérer les documents comme le fait IBM Watson dans ses outils d'ingestion. Il est par exemple capable de récupérer les résultats financiers de nombreuses entreprises et d'en présenter une synthèse. La méthode relève de la force brute au lieu d'exploiter la chimère du *web sémantique* qui n'a jamais vraiment vu le jour. Comme le web sémantique demandait un encodage spécifique et structuré des données, peu de sites l'ont adopté et l'extraction de données reste empirique. Le traitement même de ces données pour les interroger n'a pas l'air de faire partie de leur arsenal. Depuis quelques années, WorkFusion a l'air de s'être repositionné sur le secteur de la RPA.

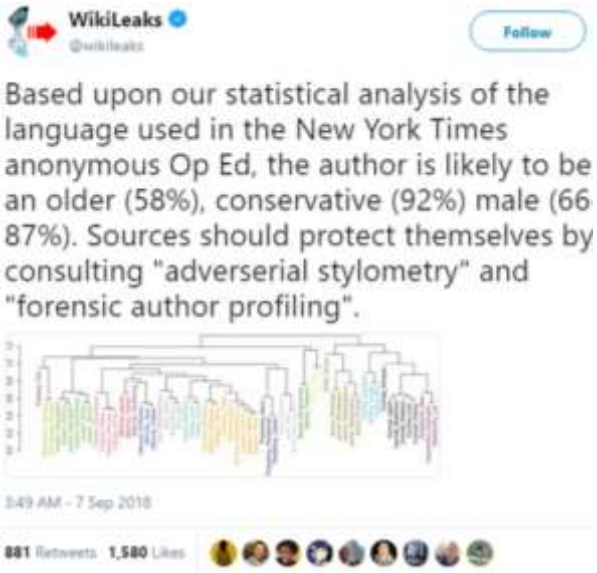
DefinedCrowd (2015, USA, $12,5M), capte les données vocales ou textuelles et les exploite, notamment pour de l'analyse de sentiments. On n'échappe pas à sa **vidéo** avec son ukulélé de circonstance.

Weotta (2011, USA) met en œuvre ce genre de technique dans son application WeottaGo, une application de recommandation mobile.

Il existe évidemment divers outils de développement spécialisés dans le traitement du langage. On peut citer notamment la bibliothèque open source **Gensim** écrite en Python qui sert notamment à analyser des textes, à identifier des sujets traités ou des sentiments et peut-être notamment exploitée dans des applications de commerce en ligne.

L'extraction de données d'un texte devrait aussi permettre en théorie d'identifier leur auteur. C'est ce qu'ont tenté plusieurs organisations pour identifier l'auteur de la fameuse tribune libre publiée en septembre 2018 dans le New York Times, provenant d'un membre de l'administration Trump critiquant sévèrement le Président. **Wikileaks** a déduit de ses analyses que l'auteur était âgé, conservateur et un Homme. Ce qui nous avance bien puisque c'est la description type d'une bonne part des gouvernements en général.

Reste à trouver de qui il s'agit sachant que l'auteur a volontairement brouillé les pistes en utilisant des tournures de différents membres de l'Administration Trump [924]!



Moteurs de recherche

Les moteurs de recherche se sont développés avant que l'IA ne devienne « mainstream » mais ils font de plus en plus appel à l'IA pour améliorer leurs fonctionnalités.

L'IA joue notamment un rôle clé dans la recherche d'images[925], pour les tagger automatiquement ou pour rechercher des images similaires. Cette dernière fonction s'appuie de plus en plus sur du deep learning et des réseaux de neurones convolutifs.

Les outils de traitement du langage naturel sont aussi mis en œuvre pour comprendre le contenu et le contexte des recherches et pour décoder la voix des vidéos.

Google utilise depuis quelques années son outil maison RankBrain pour améliorer la pertinence des recherches, en complément de l'historique PageRank. Il serait utilisé dans plus de 15% des requêtes d'utilisateurs. Le système détermine les termes qui ont un sens voisin des mots utilisés dans la recherche en fonction de son contexte.

Antidot (1999, France, $3,5M) est connu pour son moteur de recherche pour entreprises. Il propose une fonction de classification automatique de contenus ainsi que d'amélioration de la pertinence des résultats de recherche s'appuyant sur du machine learning.

Sinequa (2002, France, $5,33M) est un fournisseur de solutions de big data et d'analyse de données pour les grandes entreprises. Il fournit un moteur de recherche sémantique capable d'exploiter les données issues de nombreux progiciels (ERP, CRM, gestionnaires de contenus, etc). La société a annoncé en 2015 investir dans le machine learning pour améliorer la performance de ses solutions.

Il existe aussi de nombreux moteurs de recherche spécialisés comme pour les métiers juridiques, vus plus loin dans la partie correspondante.

---

[924] L'auteur anonyme a révélé son identité fin octobre 2020. Il s'agit de Miles Taylor, qui avait joué le rôle de chief of staff de Kirstjen Nielsen, qui dirigeait alors le Department of Homeland Security.

[925] Voir Image search using multilingual texts: a cross-modal learning approach between image and text, par les équipes de Qwant, 2019 (9 pages) qui décrit une méthode de recherche d'images multilingue.

Dans le domaine de la recherche, nous avons notamment **Iris.ai** (2015, Norvège, $422K), qui facilite la fouille documentaire de travaux scientifiques et affiche des nuages de mots clés facilitant la navigation dans les résultats. Elle ambitionne aussi d'automatiser certaines fonctions des chercheurs sur le plus long terme (vidéo). La startup organise aussi des Scithons, des hackathons scientifiques permettant la mise en réseau de chercheurs et d'entreprises.

Enfin, **Reverso** (1998, France) propose un dictionnaire des synonymes et d'analogies en ligne ou sur application mobile qui comprend plus d'un million de synonymes dans chacune des 10 langues proposées. Les synonymes sont découverts lors de phases d'entraînement en identifiant des contextes similaires dans lesquels ils sont employés. Cela complète une offre de traduction entre 13 langues différentes.

Générateur de textes et de résumés

La génération de texte soit à partir de données brutes issues de bases de données ou de résumés, soit à partir de textes est un autre pan du marché[926]. Nombre de ces solutions sont exploitées dans les médias comme nous le verrons dans la partie correspondante.

Ces logiciels s'appuient sur des modèles de texte génératifs. Ils ont des cas d'usage variés, le principal étant la génération de réponses dans des chatbots. L'objectif ultime est de créer des systèmes capables de discuter de tout sujet, histoire de se rapprocher de la réussite du fameux test de Turing de 1950. Du côté de la force obscure, ces modèles génératifs se font aussi remarquer par leur capacité à créer des fake news relativement convaincantes. Sachant qu'il en faut peu pour amadouer les fanas du genre. Ils servent aussi à résumer des textes voire à en générer à partir de bribes d'information.

Les modèles génératifs de texte les plus connus sont **BERT** et ses dérivés ainsi que **GPT**, qui nous provient d'OpenAI et dont la dernière version GPT-3 lancée mi-2020 s'est fait remarquer par la taille de son réseau de neurones comprenant 175 milliards de paramètres. GPT-3 est capable de produire de la fiction[927], des poèmes (en anglais), des communiqués de presse, des blagues, du code, des manuels techniques et même des fake news[928]. GPT-3 peut servir à prédire le futur, mais en logique très floue[929]. Il peut aussi répondre à des interviews, en étant un peu langue de bois mais pas beaucoup plus que nombre de personnes interviewées[930]. Microsoft en a acquis l'exclusivité de l'exploitation commerciale.

GPT-3 est alimenté par plus de données que n'importe quel cerveau humain pourrait en retenir : la base CommonCrawl qui contient une bonne part de l'Internet ouvert, tout Wikipedia, plus des millions de livres). C'est le système qui est le plus proche du passage du test de Turing à ce jour. L'outil soulève plein de questions aussi bien éthiques que philosophiques[931]. Recèle-t-il une conscience ? Est-il biaisé ? Si oui, il compile les biais d'une bonne partie de l'humanité, y compris des biais qui doivent être contradictoires[932]. Surtout si l'entraînement utilise tout ce qui se dit dans les réseaux sociaux et dans les commentaires d'articles !

---

[926] Sur la génération de résumés, voir A Deep Reinforced Model for Abstractive Summarization de Romain Paulus, Caiming Xiong, et Richard Socher, novembre 2017.

[927] Voir GPT-3 Creative Fiction, 2020.

[928] Voir This AI Generated Blog Made It To The Top Of Hacker News par Tyler Lee, août 2020. Encore avec GPT-3, This extraordinary AI has stunned computer scientists with its writing ability par Prasenjit Mitra, 2020 et AI-Generated Text Is the Scariest Deepfake of All par Renee DiResta, 2020.

[929] Voir An artificial intelligence predicts the future par GPT-3, 2020.

[930] Voir How I (sort of) interviewed an artificial intelligence par Tom Standage, décembre 2019.

[931] Voir Philosophers On GPT-3 (updated with replies by GPT-3) par Justin Weinberg, juillet 2020.

[932] Voir Conversational AI Can Propel Social Stereotypes par Sharone Horowit-Hendler, janvier 2020.

Le paradoxe de ce genre de système comme tous les systèmes de deep learning est que GPT-3 est immensément riche côté accumulation d'information mais qu'il ne comprend rien[933] ! Il n'a qu'une vision statistique des textes exploités, pas une vision symbolique ni de capacité à développer une logique associative ou par analogie. Il faudrait les tester sur l'informatique quantique : les entraîner avec toutes les données disponibles sur le sujet et expliquer ensuite le pourquoi du comment de la sphère de Bloch ou des états mixtes, deux sujets assez ardus.

Au nom bien choisi, **Turing-NLG** produit résumés et réponses à des questions. Créé par Microsoft Research début 2020, il comprend 17 millions de paramètres[934].

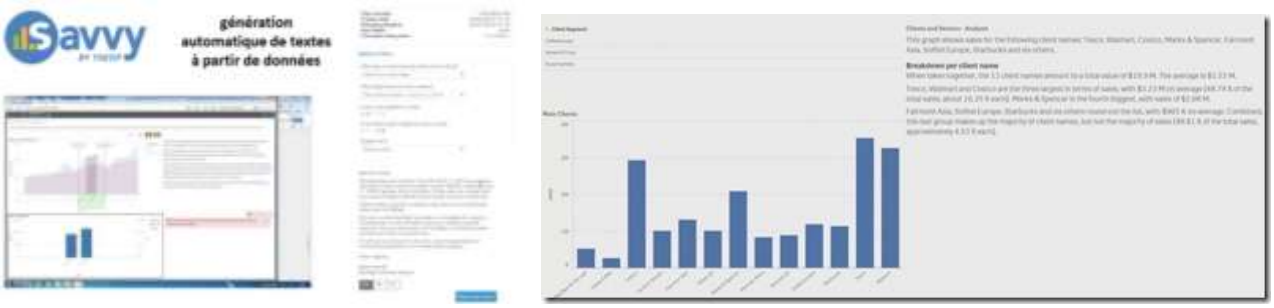
Google a de son côté développé Meena qui ambitionne de gérer des conversations sur n'importe quel sujet. Il ne comprend que 2,6 milliards de paramètres et est entraîné avec 341 Go de données textuelles[935]. Les chercheurs de **Facebook** travaillent aussi sur la question et ont notamment publié en 2020 un papier sur la création d'*open-domain chatbots*, des agents conversationnels capables de parler de tout[936].

En France, nous avons **CamemBERT** qui donne de bons résultats avec une quantité raisonnable de données d'entraînement, une base de textes de 4 Go[937]. Il peut être entraîné avec le nouveau jeu de 25 000 questions/réponses FQuAD, créé dans la lignée de Squad[938]. CamemBERT a été développé par une équipe mixte Inria, Sorbonne Université et Facebook FAIR Paris.

Le projet concurrent **FlauBERT** a été développé de son côté par une équipe associant le LIG (Grenoble), LAMSADE (UMR CNRS-Université Paris Dauphine) et LLF (CNRS). Ce dérivé de BERT est alimenté par un corpus de 71 Go de texte et comprend 137 millions de paramètres. Il a été entraîné sur 800 GPU du supercalculateur Jean Zay du centre de calcul du GENCI.

Hugging Face (2016, France/USA, $20,2M) a créé à l'origine un chatbot mobile. Ils ont pivoté en créant le modèle de traitement du langage open source Transformers qui est très populaire chez les développeurs et supporte PyTorch et TensorFlow. Il sert à extraire des données de textes, classifier des textes, faire des résumés et gérer des conversations. Résultat, la startup qui est à cheval entre Paris et New York déploie ses services pour accompagner les développeurs qui intègrent Transformers dans leurs applications. Reste à en faire une activité un tant soit peu scalable !

Le franco-américain **Yseop** (2008) est un de ces spécialistes. Basé à Lyon et à Dallas, il propose notamment Savvy, un plugin pour Excel qui traduit en texte compréhensible les données d'un graphe. Les techniques employées associent un moteur de règles et des algorithmes génétiques. Il a un concurrent américain, avec l'outil Wordsmith d'**Automated Insights** (2007).



[933] Voir AI still doesn't have the common sense to understand human language par Karen Hao, janvier 2020.

[934] Voir Turing-NLG: A 17-billion-parameter language model by Microsoft, février 2020.

[935] Voir Towards a Conversational Agent that Can Chat About…Anything par Daniel Adiwardana et al, janvier 2020 qui fait référence à Towards a Human-like Open-Domain Chatbot par Daniel Adiwardana et al, février 2020 (38 pages).

[936] Voir Recipes for building an open-domain chatbot par Stephen Roller et al, 2020 (25 pages) et FQuAD: French Question Answering Dataset par Martin d'Hoffschmidt et al, 2020 (15 pages).

[937] Voir CamemBERT: a Tasty French Language Model par Louis Martin et al, novembre 2019 (17 pages).

[938] Voir FQuAD: French Question Answering Dataset par Wacim Belblidia, mars 2020.

Narrative Science (2010, USA, $43,4M) est ainsi capable de rédiger tout seul des textes à partir de données structurées quantitatives et non structurées, avec son outil Quill. Il est utilisé dans les médias et dans le marketing. C'est un peu un équivalent des solutions du français Yseop. L'un des usages typiques est de produire une brève d'information sur le cours de la bourse ou les résultats trimestriels d'une société. Quill permet notamment d'exploiter les données issues de Google Analytics ou d'historique de transactions financières. La startup a été créée par un ancien de Google et de Carnegie Mellon.

C'est une information dont le formatage est très répétitif. La startup vise les marchés de la distribution, financiers et les services publics. La société complète depuis 2016 les textes qu'elle génère avec des graphes générés par la startup **Qlik** (1993, USA).

De nombreuses startups sont positionnées sur ce secteur, comme **Arria NLG** (2011, UK, $40,3M) qui vise les marchés financiers, des utilities, de la santé et du marketing, les français **LabSense** (2011, France) et **Syllabs** (2006, France, $2M), **Articolo** (2014, Israël) qui produit des articles automatiquement, **Retresco** (2008, Allemagne) qui produit automatiquement des comptes-rendus de compétitions sportives et **Textomatic** (2015, Allemagne) ou **Automated Insights** (2007, USA, $10,8M).

Enfin, citons côté applications grand public, le cas de **Google Smart Compose**, présenté lors de Google I/O en mai 2018, qui remplit les mails automatiquement dès que l'on tape quelques mots (vidéo).

Travail collaboratif

À des niveaux divers, les outils de travail collaboratif intègrent progressivement des briques d'IA, en particulier de traitement du langage. On trouve ces briques d'IA depuis quelques temps logées dans les systèmes d'antispams.

Dans les CMS (Content Management Systems), l'IA peut servir à analyser les textes pour détecter la langue, la tonalité, les sentiments, les profanités et le sujet. Elle peut aussi labelliser automatiquement les images partagées et générer des transcripts de vidéos en identifiant les locuteurs. L'analyse du langage et le machine learning permettent enfin de personnaliser en 1/1 les contenus diffusés.

L'IA aide le processus éditorial dans la traduction, le sous-titrage de vidéos et la création de leurs transcriptions. A terme, les CMS devraient pouvoir analyser les parcours clients et prédire la nature des contenus les plus efficaces pour faire avancer les clients dans le cycle de vente[939].

[939] Voir The Role Of AI In The Future Of Content Management Systems par Mayank Mishra, août 2019.

Contentstack (2018, USA) est un des éditeurs de CMS qui a pris le pas de l'IA pour l'intégrer dans son offre et se connecter avec IBM Watson et Salesforce Einstein[940]. Il en va de même pour **Alfresco** (2005, USA)[941] qui utilise de son côté l'IA pour faire de l'extraction de données structurées et non structurées dans les documents partagés dans l'entreprise. Ils s'appuient pour cela sur les services d'Amazon (Comprehend, Rekognition et Textract).

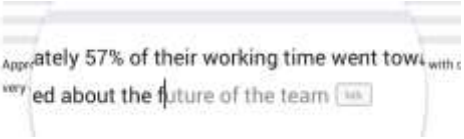
Recital.ai (2017, France) s'est attelée au problème de la gestion des mails répétitifs pour faire gagner du temps aux cadres dans les entreprises. Leur solution exploite des exemples de mails échangés dans l'entreprise. Elle classe automatiquement vos mails par type de catégories et thèmes, et identifie l'intention et les questions de l'émetteur. Le logiciel cherche la réponse dans le système d'information de l'entreprise, via des Api (SalesForce, ect) ou déclenche un processus (reporting, autre). Il propose alors une réponse et va la placer dans les brouillons. Si vous êtes OK, vous pourrez envoyer la réponse, quitte à la modifier avant.

Nudge.ai (2014, Canada, $4M) propose un plugin pour Chrome qui analyse vos emails et autres communications (limitées a priori au navigateur) pour vous proposer avec qui rentrer en relation. Cela serait adapté aux commerciaux et concurrencerait les solutions Einstein de Salesforce qui ont une finalité voisine. En fait, la solution s'intègre avec Salesforce. L'histoire ne dit pas quelle technique d'IA est exploitée dans l'outil. Probablement des briques de traitement du langage. C'est commercialisé de 0 à $60 par mois selon le niveau de fonctionnalités.

Deckrobot (2017, USA, $2,5M) propose un plug-in de Powerpoint qui utilise l'IA pour automatiser la création de vos présentations. Les fonctions principales relèvent de l'alignement et l'espacement automatique, fort approprié, d'éléments textuels et visuels dans les slides (vidéo). On est en droit cependant de se demander s'il ne s'agit pas d'IA washing car il n'est pas évident d'en trouver des traces visibles.

Microsoft Teams comprend de nombreuses fonctions faisant à du traitement du langage : retranscription automatique de réunions audio ou vidéo, sous-titrage des prises de parole, le tout en multilingue. L'assistant vocal Cortana permet de lancer un appel, de rejoindre une réunion ou d'envoyer un message par chat sans bouger le petit doigt.

De son côté, **Google** intègre depuis début 2020 la fonction Smart Compose dans Google Docs, la même que celle qui dans Gmail permet de terminer automatiquement les phrases que l'on démarre.



Et Google Meet comprend aussi une fonctionnalité de suppression de bruit de fond. Elle fonctionne plus ou moins bien et comme rien n'est parfait dans l'IA, elle modifie au passage la voix et en retire une partie des hautes fréquences[942].

Zoho (1996, Inde/USA) propose une suite bureautique en cloud qui concurrence Office 365 de Microsoft. Au-delà du classique correcteur orthographique et grammatical, s'y ajoute un test de lisibilité de type Flesch-Kincaid et des conseils d'écriture. Cela va jusqu'à la proposition de tournures plus élégantes dans les textes. La fonction tableur fait des recommandations automatiques de charts en fonction de la structure des données[943].

Enfin, dans ce domaine, **Outmind** (Paris), une startup issue de l'incubateur de l'École Polytechnique, propose un outil à base de traitement du langage permet de faire de la recherche de noms de personnes et d'informations associées dans les intranets de l'entreprise.

---

[940] Voir AI and automation are making office life easier par Andrew Tarantola, février 2019.

[941] Voir Alfresco CMS now using AI, ML to pull, organize data par Tanner Harding, avril 2019.

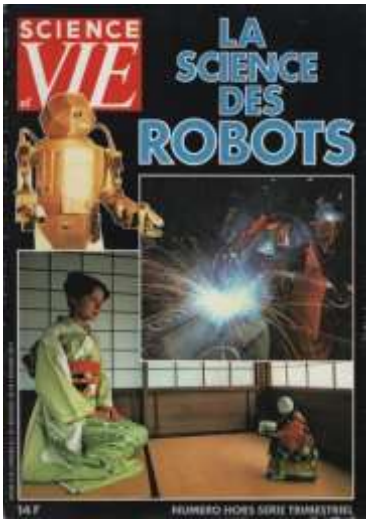
[942] Voir Google Meet's AI-Powered Noise-Cancelling Feature Is Rolling Out To Users par Tyler Lee, juin 2020.

[943] Voir Zoho s'appuie sur l'IA pour affiner sa suite bureautique par Mark Hachman et adapté par Nicolas Certes, février 2019. Zoho (Inde).

Robotique

La robotique est un domaine à part entière qui tire de plus en plus partie des briques techniques de l'IA. Au même titre que l'IA, la robotique est une science intégrative qui associe capteurs, mécanique, batteries et logiciels.

La notion de robot remonte à l'antiquité mais le mot serait apparu en 1920 dans une pièce de l'écrivain tchèque Karel Capek. Le premier robot mobile capable de s'adapter à son environnement était **Shakey** (1966-1972). Il était équipé de divers capteurs dont une caméra et des détecteurs de proximité et relié à des mini-ordinateurs DEC PDP-10 et PDP-15 via une liaison radio.



J'ai retrouvé dans mes archives un numéro spécial de Science et Vie dédié à la science des robots. Tout y est : les robots d'usines, les robots humanoïdes, les exosquelettes, etc. À ceci près que ce numéro spécial date de 1982 ! Le texte qu'il contient est toujours étonnamment d'actualité, illustrant peut-être le fait que le champ de la robotique n'avance pas si vite que cela. Un numéro spécial sur l'informatique de cette période, correspondant à l'arrivée de l'IBM PC, serait bien plus dépassé !

Classes de robots

Les définitions d'un robot ont évolué avec le temps. Aujourd'hui, on évoque un engin interagissant avec le monde physique pour accomplir diverses tâches et qui s'adapte à l'environnement.

Voici une gradation de la notion d'automate et de robot de mon cru :

- **Automate** qui répète à l'identique un geste programmé, via un logiciel ou par la saisie, d'un geste humain. C'est là que l'on peut ranger les machines d'usinage à commande numérique, les robots de peinture qui exécutent systématiquement le même geste ainsi que les imprimantes 3D. Les robots de chirurgie télécommandés sont aussi dans cette catégorie.
- **Robot qui réagit à son environnement avec des règles programmées** de manière traditionnelle par logiciel. C'est le cas d'un robot d'embouteillage qui sait s'arrêter si un incident est détecté par des capteurs simples. Les premiers robots de cette catégorie ont été créés par Unimation et installés chez General Motors en 1961. De nombreux robots industriels manipulateurs ont été créés pendant les années 1970, aux USA (Cincinnati Milacron, Unimation), au Japon (Hirata) et en Suède (ASEA).
- **Robot qui réagit à son environnement grâce à des sens** faisant appel à de l'IA et notamment à de la vision artificielle. C'est notamment le cas de nombreuses catégories de drones et de certains robots humanoïdes. Cette catégorie de robots évolue donc en liaison étroite avec les progrès récents de l'IA notamment dans l'apprentissage profond.
- **Robot doté de capacités d'apprentissage** et d'adaptation, en plus des fonctions précédentes. Ils sont plutôt rares. C'est encore un domaine en devenir, du champ de la recherche. La cobotique est située dans cette catégorie, à cheval avec la précédente. Elle comprend les robots qui collaborent avec des humains pour réaliser certaines tâches comme par exemple le transport de pièces détachées dans des machines-outils en usine.

Les robots sont souvent dédiés à des tâches dangereuses (centrales nucléaires, déminage), répétitives (peinture), stressantes (assemblage en usine), fatigantes (manutention, BTP, tonte de la pelouse), ennuyeuses (vissage), répugnantes (nettoyage) ou impossibles à réaliser de manière classique (rovers sur Mars, drones aériens).

Ils interviennent surtout là où ils sont moins chers dans la durée que des opérateurs humains et cela dépend du coût de la main d'œuvre pays par pays et du niveau d'industrialisation. Il y a par exemple très peu de robots en Afrique qui est un continent faiblement industrialisé, même au niveau de son agriculture.

La robotique nécessite l'intégration de très nombreuses disciplines : la mécanique, les moteurs, les capteurs et les sens artificiels (vision, toucher, ouïe, détection de gaz, d'humidité, de pression, de température et de proximité), la planification et le raisonnement.

Un robot est composé de très nombreux agents qui doivent être bien coordonnés. Il doit accomplir des tâches avec plus ou moins de degrés de liberté et d'initiative. Il doit pouvoir s'adapter à son environnement et gérer les imprévus. Et enfin, il doit respecter les fameuses lois de la robotique de l'écrivain Isaac Asimov issues de "I, Robot" (1950)[944].

Les sciences de la robotique cherchent à répondre à de nombreuses questions clés :

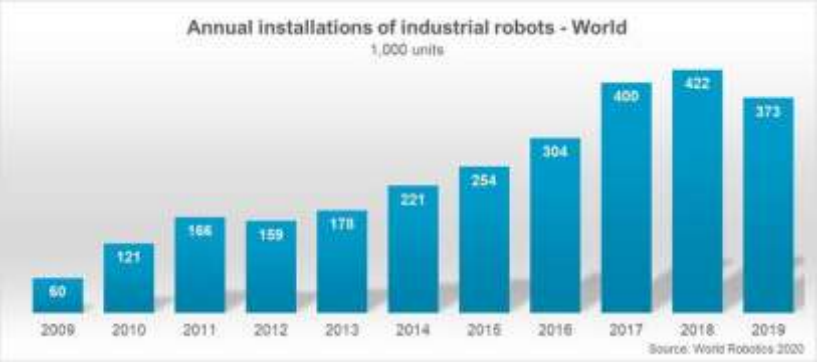
- Comment le robot peut-il se représenter le monde qui l'entoure ? C'est une question d'interprétation de ses sens visuels et autres.
- Comment doit-il réagir aux événements qu'il perçoit ?
- Comment peut-il apprendre de l'expérience ? Comme lorsqu'il apprend à éviter un obstacle de manière préventive et non pas au dernier moment.
- Comment doit-il interagir avec l'utilisateur ?
- Comment équilibrer ses objectifs et les contraintes de son environnement ?
- Comment peut-il planifier ses tâches ?

Marché des robots

En 2016, **IDC**[945] indiquait que le marché de la robotique mondial était de $71B en 2015 et devait atteindre $103B en 2018 puis $135B en 2019, générant une croissance annuelle moyenne (CAGR) de 17%. Le marché des services en robotique était estimé de son côté aux alentours de $9B à $11B en 2015 selon les sources. Fin 2018, IDC révisait ses prévisions à la baisse.

Le marché de 2019 passait à $103,4B[946]. En janvier 2020, IDC prévoyait que le marché de la robotique atteindrait $128,7B en 2020.

Il s'est vendu presque 300 000 robots industriels en 2016 et c'est la Chine qui capte la plus grande partie de ce marché, tant en production qu'en installations. Le volume était monté à 422 000 en 2018 pour baisser à 373 000 en 2019, selon World Robotics 2020 de l'IFR[947].



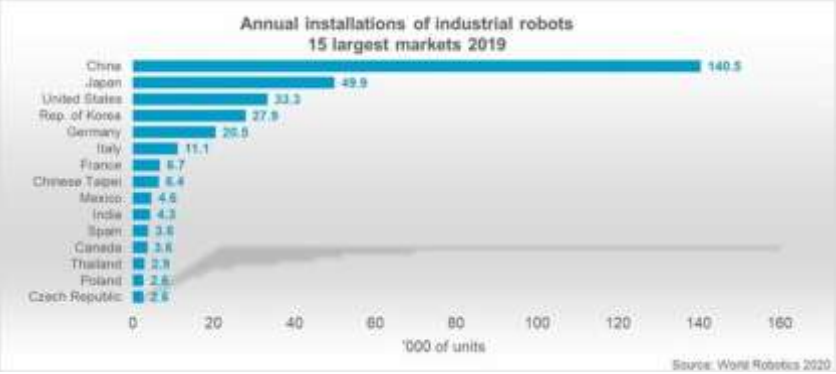
[944] Un robot ne peut porter atteinte à un être humain, ni, en restant passif, permettre qu'un être humain soit exposé au danger. Un robot doit obéir aux ordres qui lui sont donnés par un être humain, sauf si de tels ordres entrent en conflit avec la première loi, et un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.

[945] Voir The Multi-Billion Dollar Robotics Market Is About to Boom, 2016.

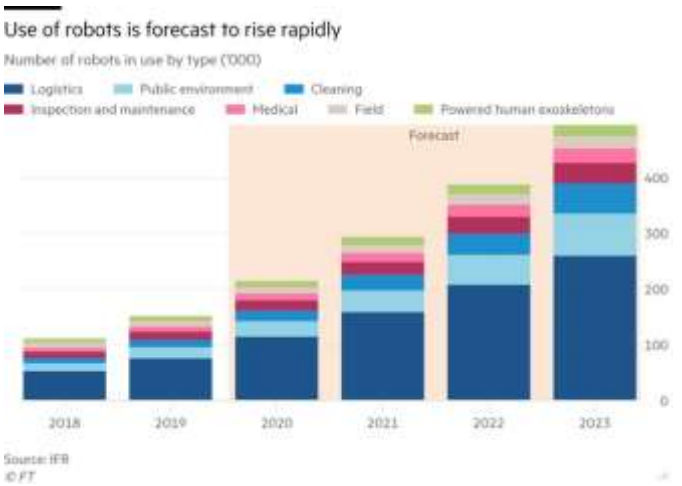
[946] Voir Worldwide Spending on Robotics Systems and Drones Forecast to Total $115.7 Billion in 2019, According to New IDC Spending Guide, décembre 2018.

[947] Voir IFR presents World Robotics Report 2020, septembre 2020.

Les gros consommateurs de robots sont sans surprise les grands pays industriels : la Chine, la Corée du Sud, le Japon, les USA, l'Allemagne et l'Italie, qui est devant la France. Le rapport de l'IFR estime que 2,7 millions de robots opéraient dans les usines dans le monde en 2019.



Le marché pourrait avoir connu un rebond en 2020 lié à la pandémie covid-19, en particulier le secteur de la robotique industrielle selon des données de l'IFR, *ci-contre*[948].



C'est la seconde source de croissance d'importations aux USA, après les produits pharmaceutiques.

Le marché est surtout concentré sur les robots industriels dans l'automobile, les industries manufacturières en général, l'agro-alimentaire et la pharmacie.

Mais ce sont les robots humanoïdes qui font le plus parler d'eux. Les japonais cherchent depuis des décennies à conçevoir de tels robots capables de s'occuper de leur population vieillissante. La population japonaise est en déclin d'environ 450 0000 personnes par an du fait d'un faible taux de natalité, situé à une moyenne de 1,43 enfant par femme. C'est aussi un choix technologique lié à un choix politique de ne pas favoriser l'immigration pouvant combler le manque de personnel soignant.

Depuis 2018, les choses ont changé : la robotique de service n'étant pas au point, et la main d'œuvre manquant, le gouvernement japonais a relativement discrètement décidé d'ouvrir prudemment les vannes de l'immigration du travail pour pouvoir prendre en charge ses seniors et couvrir les besoins de ses industries ! Il y avait déjà 1,46 million de travailleurs étrangers dans le pays en 2018. Les nouveaux travailleurs proviennent prioritairement de Chine, des Philippines et du Vietnam[949].

Robots humanoïdes

Le robot humanoïde le plus avancé du côté de sa capacité à se mouvoir est probablement Asimo de **Honda**, créé en 1986 et régulièrement mis à jour depuis. Il danse, court, monte et descend les escaliers et peut aussi tourner en rond. Sa dextérité est cependant moyenne et il n'est pas très fiable. C'est toujours un engin de laboratoire et de démonstration, dont les versions successives sont généralement construites à une douzaine d'exemplaires. Malheureusement, Honda a abandonné le projet Asimo en juin 2018 !

[948] Voir Pandemic boosts automation and robotics par Valentina Romei dans le Financial Times, octobre 2020.

[949] Voir Japan's big dilemma: Robots or immigrants? par Rebecca Townsend, mars 2019.

Cette soif de robots explique les investissements de Softbank, d'une part avec l'acquisition en 2013 du français **Aldebaran Robotics**, devenu **Softbank Robotics** et d'autre part, avec celle de **Boston Dynamics** auprès de Google en 2017[950]. Les robots humanoïdes Nao et Pepper de Softbank Robotics illustraient l'état de l'art au milieu des années 2010 avec une belle capacité de mouvement grâce à une mécanique de bon niveau, surtout pour Nao. Ils interagissent en parlant avec l'utilisateur, mais de manière encore limitée.

Pepper était censé capter les émotions des humains qu'il a en face de lui, grâce à IBM Watson, mais sa capacité de dialogue est encore approximative, limitée par les chatbots qui ont été développés pour. On n'en a pas beaucoup entendu parler depuis.

Ces robots sont des kits de développement sur lesquels sont construits des applications métiers comme par exemple un agent de renseignement pour un centre commercial, un point de vente (Softbank ou Nespresso à Tokyo) ou un lieu de transport. Leurs agents conversationnels sont des chatbots vocaux qui ne répondent qu'à des questions scriptées et en fonction des bases de données ou de connaissances auxquelles ils sont connectés.

Les robots les plus impressionnants du moment n'ont aucune capacité de dialogue. Ce sont ceux de **Boston Dynamics**, Spot Mini, Handle (*ci-dessus, à droite*) et Atlas, Handle étant capable de rouler avec habileté et de déplacer des paquets dans des étagères[951] et Atlas, de faire un saut périlleux arrière, de courir en extérieur[952] et de faire encore d'autres acrobaties régulièrement améliorées[953] et même de danser[954].

Leur capacité à comprendre leur environnement en temps réel constitue un réel progrès. C'est de l'agilité mécanique ou « athletic AI » selon le fondateur de Boston Dynamics, Marc Raibert. Elle est aussi difficile à reproduire que le raisonnement.

Ces robots sont toujours des prototypes et des démonstrateurs, pas encore des produits déployés dans les entreprises. Il en va de même pour le robot Coréen **DRC-Hubo** qui a bien du mal à se mouvoir dans l'espace de manière agile (vidéo).

---

[950] Sachant qu'en décembre 2020, on apprenait que Softbank avait cédé Boston Dynamics au Coréen **Hyundai**.

[951] Boston Dynamics faisait en 2019 l'acquisition de **Kinema Software** (2015, USA, $1,8M) une startup leur apportant des outils de vision artificielle et de robotisation dans la manipulation d'objets en entrepôts.

[952] La vidéo du saut-périlleux arrière du robot Atlas : est impressionnante mais il semblerait que ce soit un mouvement appris. Voir aussi sa course en extérieur ainsi que son agilité de coureur avec son état en octobre 2019.

[953] Voir cette autre vidéo d'Atlas de Boston Dynamics : More Parkour Atlas, septembre 2019 (38 secondes).

[954] Voir How Boston Dynamics Taught Its Robots to Dance par Evan Ackerman, janvier 2021.

Spot Mini est un robot-chien capable de manipuler des objets simples avec son bras télescopique. Les vidéos de ses démonstrations sont impressionnantes, notamment lorsqu'il ouvre une porte et qu'une personne l'en empêche et qu'il résiste[955]. Spot Mini est pour l'instant télécommandé. Son autonomie concerne la capacité à se mouvoir, pas à prendre des décisions sur la direction à prendre et le geste à accomplir ! Cela n'apparaît pas toujours dans les vidéos de démonstrations.

On peut par contre le constater dans les démonstrations publiques comme celle qui était réalisée lors de la conférence **USI 2018** à Paris (vidéo) ou dans la démonstration du Web Summit 2019 (voir la vidéo à partir de 4h46mn). Cependant, Boston Dynamics a annoncé que la commercialisation de Spot Mini avait démarré en septembre 2019. Les premiers clients auraient déjà été livrés (vidéo de lancement). Ces robots sont commercialisés en leasing. Spot Mini est notamment utilisé, équipé d'un détecteur de radioactivité, dans le démantèlement du réacteur 4 de la centrale nucléaire de Tchernobyl[956].

Boston Dynamics a déjà vendu plusieurs centaines de Spot Mini. À $75K la pièce. Il était notamment employé pour encourager la distanciation sociale dans les lieux publics à Singapour pendant la pandémie covid-19[957].

Il est notamment commercialisé par **Intuitive Robots** (2015, France, $58M), une startup créée par des anciens d'Aldebaran Robotics. Ils développent des solutions logicielles pour Spot Mini. Boston Dynamics a été acquis en 2020 par le Coréen **Hyundai** en 2020.

[955] Voir la vidéo Testing Robustness, février 2018.

[956] Voir Boston Dynamics' Spot Is Helping Chernobyl Move Towards Safe Decommissioning par Evan Ackerman, novembre 2020.

[957] Voir Boston Dynamics' Robot Dog Patrols The Streets Of Singapore To Remind People To Social Distance par Tyler Lee, mai 2020.

En 2019, des chercheurs de la **Florida Atlantic University** franchissaient la vallée de l'étrange en ajoutant d'inquiétants yeux à Spot Mini dans leur déclinaison Astro et en le rendant pilotable par la voix[958] (vidéo). Spot Mini a d'autres émules, comme le robot acrobate **Cheetah** du MIT (vidéo, *ci-contre*) ou ANYmal d'**Anybotics** (2016, Suisse) qui vise des applications d'inspection industrielle (vidéo). Enfin, vous rigolerez bien de ces parodies de **Bosstown Dynamics** réalisées en mocap et images de synthèse (vidéo 1 et vidéo 2)[959]. **Google** y travaille aussi[960].

L'un des robots qui occupent le plus le paysage médiatique depuis 2017 est sans conteste Sophia de **Hanson Robotics** (2013, Hong Kong, $21,7M). Ce robot humanoïde comprend un visage féminin animé, dans la lignée de nombreux robots japonais qui l'ont précédé.

Le robot n'était au départ qu'un buste avec un visage animé[961]. Il a ensuite gagné des jambes rudimentaires, pas encore démontrées en public. Le robot était présenté dans le Tonight Show de Jimmy Fallon en avril 2017 (vidéo), dans diverses conférences comme à la conférence Future Investment Initiative à Riyad en octobre 2017 (vidéo) à la suite de laquelle l'Arabie Saoudite lui octroyait la citoyenneté en novembre 2017[962], aux WebSummit de novembre 2017, 2018 et 2019 à Lisbonne (vidéo) ou encore au CES 2018 (vidéo) et 2019 (*ci-contre*, croisée au Venetian).

En octobre 2018, Sophia rencontrait le Président de l'Azerbaïdjan, Ilham Aliyev après avoir obtenu un visa à l'aéroport de Bakou à son arrivée[963]. A chaque fois, le robot est démontré au niveau de ses

---

[958] Voir This Boston Dynamics-Esque Bot Has Horrifying, Human-Like Eyes It looks like it's wearing the eyes of its last victim par Jacob Banas, août 2019.

[959] Voir aussi la vidéo de making of qui montre comment la première vidéo a été réalisée avec des techniques de trucage issues du cinéma. Le mocap, ou motion capture, consiste à saisir les mouvements d'un humain en général habillé en vert avec des points de repère, qui servent ensuite à y plaquer une image de synthèse photoréaliste en 3D. Ces vidéos ont été crées par le studio Corridor Digital, basé à Los Angeles depuis 2010.

[960] Voir Exploring Nature-Inspired Robot Agility par Xue Bin de Google Research, avril 2020 qui décrit les méthodes biomimétiques servant à piloter ces robots chiens. Il est intéressant de découvrir ces travaux alors que Google a revendu Boston Dynamics à Softbank en 2017.

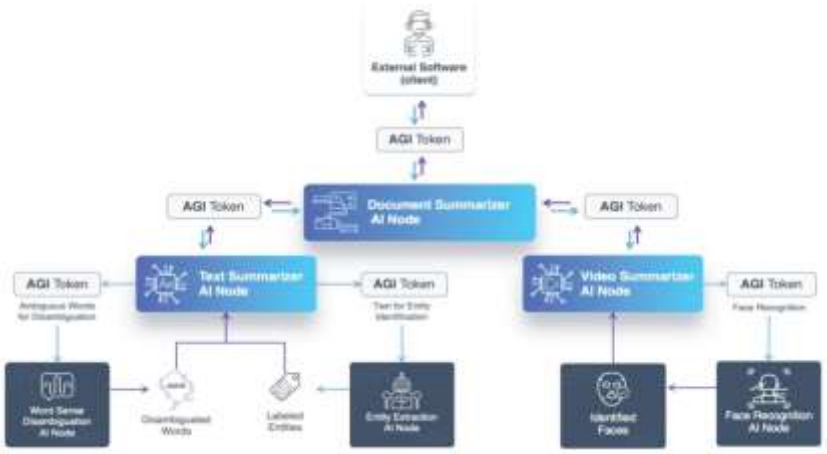
[961] A noter que la startup **Realbotix** (USA) est spécialisée dans la création de visages animés qui semblent très voisins de celui de Sophie de Hanson Robotics. Ils les installent dans des poupées féminines servant de sex-toys.

[962] Un robot citoyen ? Alors qu'il n'était même pas mobile au moment de l'octroi de cette nationalité. C'est du pipeau politique dans un pays qui n'accorde que peu de droits civiques aux ouvriers étrangers qui travaillent chez eux et qui sont souvent maltraités. Ils sont plus de 9 millions dont seulement 100 000 sont occidentaux ! "*Saudi Arabia requires foreign workers to have their sponsor's permission to enter and leave the country, and denies exit to those with work disputes pending in court. Sponsors generally confiscate passports while workers are in the country; sometimes employers also hold passports of workers' family members*" (**Wikipedia**). Le pays a autorisé les femmes à conduire, mais seulement après avoir accordé la nationalité au robot Sophia !

[963] Voir AI Humanoid 'Sophia' Is Granted First Ever Robot Visa, Speaks With President, de Paul Amstrong, octobre 2018. La vidéo du dialogue avec le Président est surtout un long monologue de ce dernier. Le robot écoute poliment et répond. Mais la voix off Azéri couvre la voix en anglais de Sophia, il est donc impossible de saisir quoi que ce soit du dialogue.

capacités d'agent conversationnel, capable de dialoguer avec son interlocuteur sur n'importe quel sujet. Mais toutes ces démonstrations relèvent de l'esbroufe et sont soigneusement préparées à l'avance. Et une bonne partie du public gobe sans broncher ! Cela fait râler à juste titre les spécialistes de la robotique en France.

Les conversations de Sophia sont en grande partie prescriptées, et ne sont donc pas spontanées [964] ! L'agent conversationnel de la startup doit s'appuyer sur la solution à base de blockchain Ethereum de l'initiative **SingularityNET** lancée par Ben Goertzel, le directeur scientifique de Hanson Robotics, par ailleurs un des parangons de l'intelligence artificielle générale et ardent promoteur de la singularité.



Sophia est censée à terme s'appuyer sur **SingularityNET**, une initiative croisée de l'**OpenCog Foundation** (une autre création de Ben Goertzel, qui travaille aussi pour Hanson Robotics), de **Vulpem** (une société de conseil spécialisée en Blockchain) et de **Novamente** (société de services dans l'IA). Le framework open source **Opencog** comprend des briques d'IA symbolique, connexionniste et génétique. Bref, si les démonstrations de Sophia doivent laisser songeur et suspicieux, il n'en reste pas moins intéressant de surveiller de près la manière dont OpenCog pourrait évoluer dans le futur.

SingularityNET souhaite rendre interopérables les grandes briques de l'IA. C'est un projet open source créé en réaction à la tentation de contrôle de l'IA par de grands acteurs, avec probablement les GAFA en ligne de mire[965].

Cela passe par un réseau d'agents pour créer une *swarm intelligence*. C'est un système décentralisé, un peu comme un DAO de Blockchain (Decentralized Autonomous Organization). Elle s'appuie sur des smart contracts bâtis sur une Blockchain Ethereum avec la volonté de supporter d'autres blockchains à terme. La solution comprend une place de marché de briques d'IA avec un système de paiement intégré et de notation. Techniquement, ils utilisent une approche neurale-symbolique. Les usages ? Assez classiques : dans la vision, le traitement du langage et la robotique. Ben Goertzel mesure le niveau de conscience de ses créations telles que Sophia avec un indice dénommé **Tononi Phi**[966].

Sophia impressionne par le réalisme de son visage. Il est recouvert d'une peau réaliste qui est chauffée à température corporelle pour accentuer son réalisme. Au Web Summit 2019, Sophia était accompagnée d'un autre avatar robot, de l'auteur Philip K. Dick (vidéo) mais sans progrès apparent même si David Hanson et Ben Goertzel les faisaient dialoguer l'un avec l'autre. Toujours avec des interventions en partie préparées à l'avance.

---

[964] Que sait réellement faire Sophia, le robot dont l'intelligence est contestée ?, de Morgane Tual, février 2018, qui reprend les propos agacés de Yann Le Cun sur le brouhaha marketing autour de Sophia. Et la vidéo Hanson Robotics Sophia is a Fake!, janvier 2018, bien complétée par Mama Mia It's Sophia: A Show Robot Or Dangerous Platform To Mislead?, de Noel Sharkey, novembre 2018.

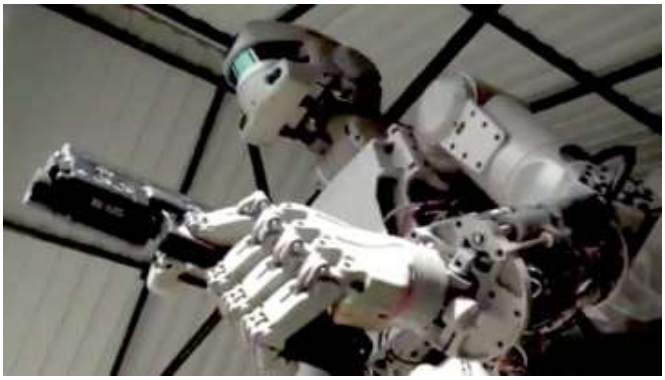
[965] Voir SingularityNET White Paper 2.0 (86 pages). Beta V2 en 2019.

[966] Voir Using Tononi Phi to Measure Consciousness of a Cognitive System While Reading and Conversing ? par Ben Goertzel & al (6 pages).

Mais ce ne sont pas les premiers robots à jouer dans cette cour. Le plus connu du genre est le **Geminoid** du Japonais Hiroshi Ishiguro (*ci-dessous à gauche)*[967]. Et en 2019, **Mark Robotic Labs** créait sa propre réplique robotisée de l'actrice Scarlett Johanson (vidéo). C'est en fait une société qui commercialise des pièces de création de robots humanoïdes. D'autres n'hésitent pas à créer de véritables faux robots, comme en Russie[968]. En 2014, le robot **Bina48** était encore plus impressionnant que Sophia. Il était même utilisé pour donner des cours à l'Université Notre Dame de Namur en Californie (vidéo) ! Il a en fait été créé par David Hanson, le fondateur de Hanson Robotics à l'origine de Sophia. Il y a aussi le double de **Nadia Magnenat-Thalman** (*ci-dessous à droite*), sans compter les nombreux animatronics qui servent aux productions cinématographiques.

Certains imaginent déjà que l'on pourra un jour tomber amoureux de tels robots, notamment là où la solitude sévit le plus comme pour les célibataires au Japon, qui n'ont pas de difficulté à projeter des émotions sur des robots ou des avatars[969]. Cela passera aussi par les robots sexuels que certains cherchent à développer, cela étant avant tout une affaire de mécanique et de matériaux innovants. À vous de voir si ce monde est souhaitable !

Ces concepts de robots humanoïdes sont relativement inoffensifs. Jusqu'au moment où l'on en crée des versions à usage militaire. C'est le cas de ce robot russe expérimental, le **FEDOR**, tout droit sorti de Robocop qui tire avec précision sur une cible fixe (vidéo). Si l'objectif est de faire peur sur les risques de l'IA et de la robotique, il sera rapidement atteint. En fait, une version civile, sans le flingue, doit rejoindre l'ISS en 2020 pour aider les spationautes à mener des tâches diverses.



De nombreux robots avec des capacités mécaniques plus limitées sont proposés pour servir de centres de renseignement ambulants dans des lieux publics comme les centres commerciaux ou les aéroports. Ce sont en quelque sorte des tablettes à roulettes, comme chez le chinois **Qihan**, le français **Hease Robotics** (2016, France), un autre robot français, créé à Lyon et animé par les logiciels d'un autre Lyonnais, la startup **Hoomano** (2014, France). Hease Robotics a fermé boutique fin 2019.

---

[967] Voir Super realistic robots at Tokyo Game Show leave social media divided with a hot debate, juillet 2018.

[968] Voir Russie : le robot high-tech dissimulait un être humain, dans Le Point, décembre 2018.

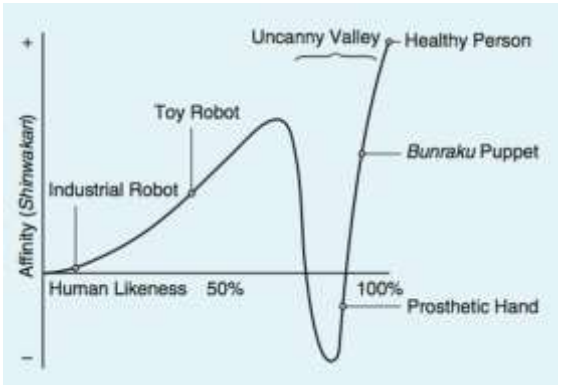
[969] Voir "Il sera un jour possible de tomber amoureux d'un robot" par Matthieu Delacharlery avec Laurence Devillers, février 2020.

L'un des ancêtres de ces robots est le **PR2 Willow Garage** de Stanford (vidéo). Créé à la fin des années 2000, il comprend deux bras articulés. C'est un kit de développement pour les chercheurs. Il se déplace grâce à des roulettes et peut bouger son corps verticalement. Il est commercialisé aux alentours de $300K.

Il y a aussi ceux d'autres startups françaises, **Partnering Robotics** (2007, France) et son Diya One et **Wyca Robotics** (2015, France) et sa base modulaire Elodie ou **Robosoft Technologies** (1996, Japon, $15,7M) et son robot compagnon Kompaï 2.

Et il y a aussi **Temi** (2015, USA, $36M) qui se présente comme étant la première société au monde fabricant un « AI robot ». En pratique, c'est une tablette à roulette de plus, pilotable entre autres choses par Amazon Alexa (*ci-contre*). Elle sert notamment à déclencher des visioconférences n'importe où dans les bureaux ou les hôpitaux.

La fonction principale de ces robots de services est l'agent conversationnel qu'ils intègrent et qui gère un spectre étroit de discussions. Je me moque un peu de ces tablettes à roulettes mais leur forme a une utilité. Elle est d'abord plus facile à fabriquer et gérer et elle permet d'éviter de faire basculer le robot dans la vallée de l'étrange (*uncanny valley*). Elle correspond au sentiment ressenti qui peut être désagréable lorsque l'on est face à un robot proche d'un être vivant mais pas parfaitement. Même si certains n'hésitent pas à franchir cette barrière symbolique[970].



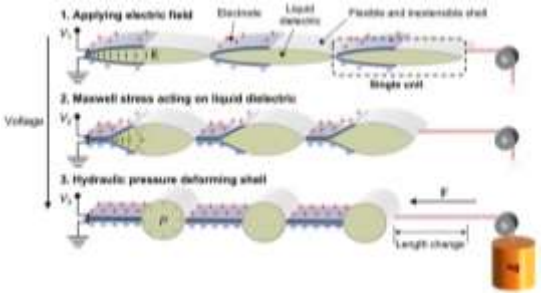
Les progrès de la robotique continuent cependant inexorablement et pas à pas.

[970] Voir Affetto : le robot enfant aux expressions tellement réalistes qu'il fait flipper par Pierrick Labbe, 2018. Le concept de la vallée de l'étrange a été proposé par le roboticien japonais Masahiro Mori. Voir The Uncanny Valley: The Original Essay by Masahiro Mori, juin 2021, qui traduit en anglais le texte original datant de 1970. Voir aussi Robophobie : faut-il avoir peur des humanoïdes ? avec Marjorie Paillon et Guillaume Grallet, octobre 2019 (12 mn) qui met bien en perspective le sujet et en images.

Avec l'amélioration du modèle prédictif du toucher, via des travaux du CSAIL du **MIT** (Computer Science and Artificial Intelligence Laboratory)[971], le développement de bras articulés[972], la prévision du comportement des humains[973], la capacité à grimper sur des murs métalliques qui est plus une prouesse mécanique qu'autre chose [974], de l'apprentissage par renforcement pour apprendre à manipuler un Rubik's Cube[975], et puis également tout ce qui touche aux processeurs permettant aux robots d'interpréter leur environnement[976].

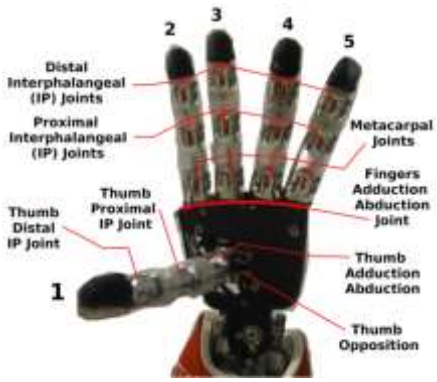
Au-delà de l'IA, il reste de gros points à traiter en robotique comme la création de batteries efficaces et de moteurs électriques adaptés. Les premières ont une densité énergétique encore trop faible, en tout cas lorsqu'elles sont basées sur le bon vieux procédé lithium-ion et les seconds n'arrivent pas à égaler la souplesse et l'efficacité énergétique des muscles striés des mammifères. D'où les intéressantes recherches visant à créer des muscles artificiels avec la même souplesse et efficacité énergétique.

Le **Keplinger Research Group** de l'Université du Colorado à Boulder est l'un des laboratoires qui travaille sur des muscles artificiels, dénommés HASEL[977] élaborés avec des élastomères diélectriques. Ils pourraient servir aussi bien en robotique que pour créer des prothèses destinées à des humains. Mais c'est encore loin de l'industrialisation.



Autre idée bio-inspirée, permettre aux robots de transpirer pour leur permettre de réguler leur température intérieure. C'est un projet de recherche de l'Université Cornell et de l'Institut Italien de Technologie qui ne concerne à ce stade qu'une main robotisée exploitant des doigts imprimés en 3D en stéréolithographie avec un hydrogel capable de rétracter des pores pour laisser évaporer de l'eau. En plus d'une recharge de batterie, il faudra donc donner à boire au robot[978] ! On aura tout vu !

Il serait intéressant de pouvoir coupler ce système avec la main articulée de **Shadow Robot Company** (1997, UK). Celle-ci est capable notamment de manipuler des objets fragiles (vidéo). Leur main Shadow Dexterous Hand est dotée de 20 degrés de liberté et de capteurs de pression tactile. Elle pèse 4,3 kg et peut soulever une charge de 4 kg si elle est bien arrimée. La main ne fait rien seule. Pour pouvoir être guidée, elle doit être complétée de briques logicielles, notamment de vision artificielle, et d'un système de décision associé.



---

[971] Voir MIT's new AI for robots can 'feel' an object just by seeing it par Ivan Mehta, 2019.

[972] Voir par exemple Reachy, a 3D-Printed Human-Like Robotic Arm as a Testbed for Human-Robot Control Strategies par Sébastien Mick et al, août 2019.

[973] Voir MIT algorithm helps robots guess where humans are going next par Jon FIngas, juin 2019. Ainsi que L'iA du MIT permet aux robots de prévoir les mouvements des humains par Denis, Didier, Ergo et Guillaume, 2019.

[974] Voir This robot climbs walls so that humans don't have to par Mark Stetson, 2019.

[975] Voir AI-driven robot hand spent hundred years teaching itself to rotate cube par Will Knight, 2019.

[976] Voir Efficient Computing for Autonomous Navigation of Miniaturized Robots, par Vivienne Sze du MIT, mars 2019 (22 minutes) qui décrit l'état de l'art des processeurs dédiés aux robots (19 slides).

[977] Voir The artificial muscles that will power robots of the future par Christoph Keplinger, TEDx, mai 2019 (11 minutes) ainsi que Peano-HASEL actuators: Muscle-mimetic, electrohydraulic transducers that linearly contract on activation par Nicholas Kellaris & Al, janvier 2018 (11 pages).

[978] Voir Robots learn to sweat to stop overheating par JamesVincent, janvier 2020.

Robots de services

De nombreux robots que nous venons de voir sont inspirés d'œuvres de science-fiction. Les robots commercialisés en volume et véritablement opérationnels prennent bien d'autres formes.

On trouve d'abord des robots de surveillance et de visio-conférence comme ceux de **Gostai**, une startup française acquise par **Aldebaran Robotics** juste avant qu'elle soit elle-même absorbée par **Softbank Robotics**.

De même, **Cobalt Robotics** (2016, USA $53,3M) propose un robot de surveillance de bureaux qui peut aussi servir à gérer des visioconférences à distance. Il détecte les incidents et alerte alors une centrale d'alarme[979].

Suivent de nombreux robots « de sol », qu'il s'agisse d'aspirateurs (comme le Roomba de **iRobot,** avec des prix compris entre 300€ et 1000€), de tondeuses à gazon (**Friendly Robotics, Husqvarna** ou chez le Français **InfinyIA**[980]) ou de systèmes de sécurité mobile (**EOS Innovation**[981], **AI.mergence**), leur principale fonction est de se mouvoir au sol, d'éviter les obstacles, de balayer optimalement une surface donnée, et de réaliser une tâche mécanique spécifique.

Leurs capteurs de proximité voire visuels leur permettent de cartographier leur environnement et de s'y mouvoir. Ils doivent aussi souvent pouvoir retrouver leur station de recharge de batterie. Une variante de ce genre de produit est l'étonnante valise robot de **Cowarobot** qui suit automatiquement son utilisateur que j'avais découverte à l'occasion du CES 2017 et qui est commercialisée depuis (vidéo). On espère qu'elle hurle si quelqu'un cherche à la voler à son propriétaire !

Les loisirs sont un autre domaine pour les robots, avec le robot joueur de ping pong du fabricant de composants **Omron** (*ci-dessous*, vu au CEATEC de Tokyo en octobre 2014 et qui n'est qu'un démonstrateur pour un fabricant de composants et qui s'améliore régulièrement) ou le ramasseur de balles de tennis de **Tennibot** (2016, USA).

---

[979] Voir Meet the autonomous security robot that will soon patrol your offices par Danica D'Souza, 2019.

[980] InfinyIA est une filiale d'Easy Center Corp Consulting, une société de conseil en administration de systèmes Microsoft. Elle propose son robot tondeuse I.MOWER T2000 présenté comme le plus intelligent du marché. Il est équipé d'un GPU Xavier de Nvisia et d'une vingtaine de capteurs dont un LIDAR rotatif et quatre caméras 360°. Ses capteurs permettent de cartographier le terrain au niveau de la qualité de la terre et la nature des végétaux rencontrés. Il est associé à une remorque elle-même motorisée. Tout cela permet à la tondeuse d'éviter les obstacles, et en bonus, les enfants. Vu son équipement matériel, c'est une Rolls destinées à des clients C+++. Son prix est de 3K€ HT (vidéo).

[981] L'ESN **Econocom** a décliné le robot d'EOS Innovation pour en faire un robot d'inspection de data-center en 2017, Captain DC. EOS Innovation était une filiale du français Parrot qui a été ensuite acquise par **Econocom**. Voir https://vimeo.com/170005575.

Le joueur de ping-pong robotisé exploite surtout un système de vision stéréoscopique couplé à un système prédictif de la position de la balle en fonction des gestes de son compétiteur humain. Il existe des robots joueurs de ping-pong avec des bras articulés 6 axes de **Kuka** (vidéo). Le ramasseur de balles de tennis utilise ses capteurs de vision pour détecter les balles et les ramasser.

Omron (2014)

Tennibot (2017)

Bionic Bar (2015)

Denso Barrista Robot (2017)

Les robots d'usines sont de leur côté mis à contribution pour devenir barmen ou gestionnaires de machine à café comme avec le Bionic Bar, installé dans les paquebots du croisiériste américain **Royal Carribean** et le **Denso Barrista Robot**, une déclinaison d'un robot d'usine pour un usage grand public de démonstration, Denso étant surtout un équipementier pour l'industrie automobile.

Il existe aussi plein de robots transporteurs de charges pour les entrepots, comme les robots manutentionnaires de **Kiva Systems** (2003, USA, $17M) qui ont été acquis par Amazon pour $775M en 2012 (*ci-dessous à gauche*).

Alibaba a des investissements qui vont dans le même sens pour automatiser ses entrepôts avec les robots de **Quicktron Robotics** (Chine, $129M) (vidéo). Ces derniers sont concurrencés par une société sino-américaine, **Prime Robotics** (2001, Chine), anciennement Bleum Robotics.

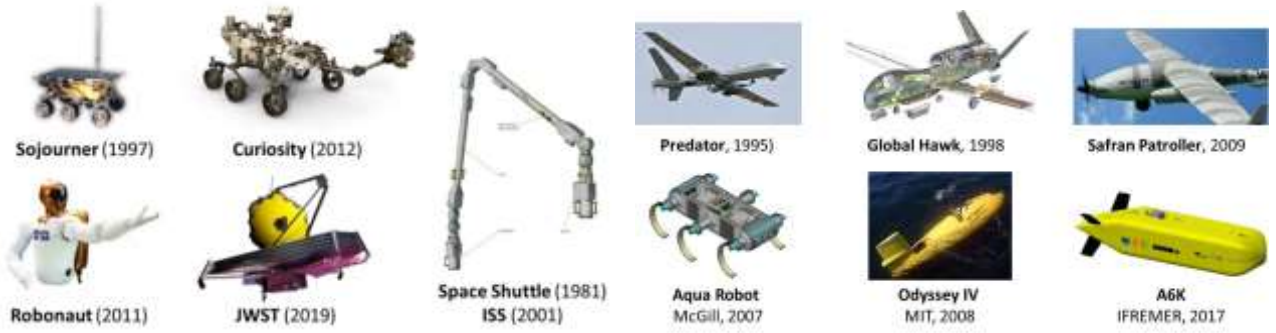
Dans les transports, les robots de **FuelMatics** (2010, Suède) remplissent automatiquement votre réservoir d'essence si vous avez installé leur bouchon spécifique dans votre véhicule (*ci-dessous à droite*).

Dans l'aérospatial, les robots doivent être très autonomes. C'est le cas des rovers **Sojourner** et **Curiosity** qui explorent Mars. Les télécommunications aller et retour entre Mars et la Terre durent plus de 45 mn. Ces robots doivent donc se débrouiller tous seuls en fonction de leur plan de charge. La conséquence est qu'ils sont plutôt lents.

FuelMatics (2008)

Les télescopes spatiaux sont aussi très autonomes, comme le **James Webb Telescope** dont les retards successifs du lancement sont désespérants. Il mettra plus de deux semaines à se déployer avec des dizaines d'opérations de dépliement de sa structure en origami.

Les drones militaires savent gérer leur vol de manière autonome mais sont pilotés à distance, notamment pour les ordres d'observations ou de lancements de missiles. Il en va de même pour les drones sous-marins qui sont téléguidés.



Au-delà des drones classiques, des chercheurs poussent le concept assez loin avec des drones capables de rattraper des balles à la volée grâce à une propulsion omnidirectionnelle à huit hélices[982], ou via la collaboration entre plusieurs drones, proche de la notion de système multi-agents[983]. On peut y ajouter cette chaise robotisée du MIT qui, en 2012, se cassait et se réparait toute seule[984].

Comment fonctionnent tous ces robots ? Un peu comme pour le Machine Learning et le Deep Learning, leurs créateurs ont rarement développé des logiciels idoines pour leur donner vie.

Ils s'appuient le plus souvent sur des SDK du marché. Plusieurs startups sont présentes sur ce marché et notamment **Vicarious** (2010, USA, $122M), **Kindred** (2013, USA, $44M), **Osaro** (2015, USA, $29,3M) et **Brain Corp** (2009, USA, $161M). Des startups comme l'Américaine **Neurala** (2006, USA, $20,1M) sont spécialisées dans l'IA pour le pilotage automatique des drones avec leur SDK Brain4Bots intégrant deep learning, vision artificielle et support de GPU comme ceux de Nvidia.

On peut aussi compter sur **Kinema Systems** (2016, USA) et son système de vision 3D pour robots de manutention, Kinema Pick associé au capteur 2D/3D (vidéo) ainsi que **RightHand Robotics** (2014, USA, $11,3M) qui est sur un créneau voisin avec des solutions robotiques intégrées pour la préparation de colis dans le commerce en ligne.

Conscience Robotics (2017, France) propose aussi via son Conscience Store des briques logicielles permettant de créer un robot ad-hoc. Leur marketing force un peu le trait en présentant cela comme une « intelligence artificielle universelle ».

C'est lié au fait que leurs briques logicielles sont exploitables pour divers types de robots : bras robotisés, drones volants, robots humanoïdes ou véhicule à roues. Il faut se garder de survendre la robotique et ses capacités d'IA[985].

Covariant (2018, USA, $67M, anciennement Embodied Intelligence) est issue de Berkeley[986] et développe des fonctionnalités permettant à des robots d'être plus versatiles dans leur gestuelle et maîtrise de la manipulation d'objets.

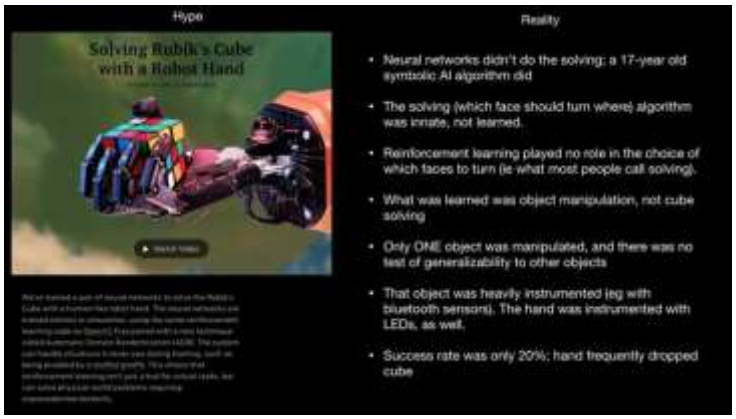
---

[982] Voir Fetching Omnicopter, 2017.

[983] Voir Cooperative Quadrocopter Ball Throwing and Catching ETH Zurich, 2012.

[984] Voir Robotic Self Healing Chair, 2017 (4 mn) et The Robotic Chair by Max Dean & Raffaello D'Andrea, 2012 (4 mn). Vient du MIT.

[985] Voir The robotic fallacy par Tim Kindberg, octobre 2019 qui rappelle que l'intelligence artificielle reste très artificielle.



Plus de vingt ans après la renaissance des réseaux neuronaux, en 2006, le japonais **Osamu Hasegawa** créait les réseaux neuronaux auto-organisés incrémentalement ("Self-Organising Incremental Neural Network" ou SOINN), utilisables dans des réseaux neuronaux auto-réplicables et capables d'auto-apprentissage.

En 2011, son équipe développait un robot utilisant ces SOINN capable d'auto-apprentissage (vidéo, *ci-dessus à droite*), illustrant magistralement les applications des réseaux neuronaux. Depuis, ces capacités d'auto-apprentissage de la préhension d'objets ont été intégrées dans un grand nombre de robots[987]. À chaque fois cependant, les tâches auto-apprises sont spécifiques et ont été préconfigurées.

La robotique est encore un grand champ d'expérimentation et de makers. Un grand nombre de projets d'étudiants et de chercheurs tâtonnent pour faire avancer le domaine. On a par exemple des robots de tri de pièces de Lego[988] ou de résolution du Rubik's Cube en 0,88s (vidéo) puis en 0,38 s (vidéo) puis avec une seule main robotisée (vidéo), une performance d'OpenAI un peu survendue[989] (*ci-dessus à gauche*).

Divers autres robots originaux qui n'ont pas forcément vocation à être industrialisés et commercialisés. Vous avez ce projet de robot menuisier domestique développé par le **MIT** (vidéo, février 2018), un robot abeille de pollinisation développé au Japon par le National Institute of Advanced Industrial Science and Technology[990] ou le robot **Salto-1P** de Berkeley qui peut bondir pour traverser une pièce (vidéo)[991].

De nombreuses startups se lancent avec plus ou moins de bonheur dans le domaine. Les me-too sont légion et on attend toujours des robots capables de bien appréhender leur environnement et d'interagir avec. Se développe également une économie de services et d'ingénierie en robotique[992].

---

[986] AI Startup Embodied Intelligence Wants Robots to Learn From Humans in Virtual Reality, novembre 2017. Voir la vidéo où leur robot peut faire des nœuds avec des ficelles, après avoir réussi le pliage de serviette (vidéo). Ils développent des techniques d'apprentissage des gestes par la réalité virtuelle (vidéo).

[987] Voir par exemple Showing robots how to do your chores par Rob Matheson, mars 2020 ainsi que Des chercheurs apprennent à un robot comment se réparer tout seul par Valentin Cimino, 2019.

[988] Voir How I Built an AI to Sort 2 Tons of Lego Pieces par Jacques Mattheij, juin 2017. Ce dernier utilisait un GPU pour réaliser le tri visuel. En 2019, un bricoleur réalisait la même performance avec un simple Raspberry Pi. Voir Cette machine trie les pièces de Lego mieux que vous ne le saurez jamais par Sitraka R, 2019 (vidéo).

[989] L'explication provient d'un tweet de Gary Marcus. Voici des détails relativisant la performance : Non, une IA robotique à une main n'a pas résolu le Rubik's Cube par Aymeric Poulain Maubant, octobre 2019.

[990] Voir Japan has just invented Robo-bees that can legitimately pollinate the earth, février 2017.

[991] Voir Insolite : Le robot le plus petit au monde qui fait des sauts de géant, de Loïc Bremme, juillet 2018.

[992] Comme avec **iSee Automation**, une startup issue du MIT qui est financée dans le cadre du fonds de deep techs **The Engine** lancé par ce dernier et doté de $200M (source).

Cela explique pourquoi nombre de métiers manuels sont bien plus protégés aujourd'hui que le sont certains métiers de cols blancs gérant des processus répétitifs, bien plus faciles à automatiser.

La robotique est une discipline complexe, à l'instar de celui des objets connectés. C'est un défi rarement relevé que d'être bon à la fois dans le matériel et le logiciel. Les sociétés développant la partie matérielle des robots sont souvent déficientes en logiciel où elles n'investissent pas assez et où l'état de l'art n'évolue pas assez rapidement. Ou alors, les robots ne sont que des tablettes roulantes qui vont s'appuyer ensuite sur les grandes plateformes de chatbots vocaux pour dialoguer avec des utilisateurs. Il y a évidemment des échecs commerciaux dans ce secteur[993].

D'où le fait que nombre de robots commercialisés sont très spécialisés comme dans l'agriculture ou dans les usines ou pour la manutention dans les entrepôts. Le robot généraliste est un projet très long terme ! C'est au minimum un gros SDK qui doit donner lieu au développement de nombreuses IA étroites pour son ou ses usages.

Ces écosystème sont lent à créer et comme dans les smartphones, il n'y a pas de place pour un très grand nombre d'acteurs. Soit le marché est fragmenté est les écosystèmes ne se développent pas, soit il se consolide au prix de l'élimination d'une bonne part des petits acteurs. Dans ce cadre, il était intéressant de visiter le stand « d'écosystème » de Pepper chez Softbank Robotics à VivaTech.

Les Français qui se lancent dans la création de startups en robotique tombent donc dans l'un des deux cas de figure. Soit des robots spécialisés, soit des robots généralistes, difficiles à mettre au point.

TwinswHeel (2016, France) crée des robots autonomes de livraisons dédiés aux environnements fermés comme des usines, semi-fermés et ouverts (vidéo). Ils ont déjà exposé au CES et été soutenus par la Poste. C'est le modèle de livraison du « last mile », le point clé étant ensuite celui des derniers mètres pour arriver entre les mains du destinataire ! Le robot qui a une charge utile de 40 kg sait bien entendu éviter les obstacles le long de son trajet. Il sait aussi traverser la rue aux passages cloutés et interpréter la signalisation. Il roule entre 6 km sur le trottoir et 30 km/h sur la route[994].

Cybedroid (2011, France) et son robot Alice qui rappelait Pepper, mais sans l'écran. Il a des mains et se déplace sur roulettes. Ses moteurs lui procurent 20 degrés de liberté et il dispose de huit heures d'autonomie pour déambuler dans des lieux publics et servir de concierge d'entreprises et accueil. La valeur est dans l'agent conversationnel métier qui motorise le robot ! Et plouf, liquidation judiciaire en novembre 2019. Une de plus dans un marché et un domaine bien difficile et balbutiant. La plateforme robotique générique reste à inventer !

[993] En 2019, **Anki** (2010, USA, $182M), le fabricant du petit robot jouet Cozmo, fermait boutique en avril 2019.

[994] La startup n'est évidemment pas seule sur ce marché qui intéresse aussi les grands du ecommerce. Voir par exemple Alibaba mise sur un robot autonome pour les livraisons, de Pierrick Labbé. Leur G Plus roule à 15 km/h mais est positionné pour un usage dans les entrepots. Dans le même ordre d'idée, voir OpenAI Demonstrates Complex Manipulation Transfer from Simulation to Real World, de Evan Ackerman, juillet 2018, avec de l'apprentissage de geste par tâtonnement et injection d'aléatoire. Le robot de Twinswheel était aussi testé fin 2018/début 2019 par Franprix pour les livraisons de proximité. Voir Franprix va tester un robot livreur de courses à Paris en 2019 par Pierrick Labbé, 2018.

Spoon (2015, France) propose un robot manipulateur d'objets ou d'écrans (vidéo). Se présentant comme un robot empathique, il permet des interactions avec le public, avec une souplesse lui donnant un côté « animal ». Le reste est dans l'écran, comme c'est le cas dans Pepper, et on retombe dans le défi logiciel et des agents conversationnels. Ses fondateurs sont issus d'Aldebaran Robotics et avaient contribué à la création de Nao et Pepper.

Ils se sont repositionnés dans le logiciel et les interactions hommes/machines, dont l'ajout d'avatars 3D à des chatbots.

Cobots

Les robots collaboratifs sont utilisés à proximité d'opérateurs pour accompagner leurs gestes grâce à de meilleures facultés d'adaptation et une bonne gestion des contraintes de sécurité. Les principaux fabricants de cobots sont à la fois des entreprises récentes comme **Universal Robots** (2005, Danemark, acquis par Teradyne en 2015), **Rethink Robotics** (déjà cité), **Micropsi** (2014, Allemagne, $9,5M, avec son robot Mirai), des fabricants industriels comme **Kuka** (qui appartient maintenant au Chinois **Midea**), **ABB**, **Fanuc** tout comme quelques startups françaises telles que **MIP Robotics** (2015, France, $2,3M) et ses robots de manutention de petites pièces et **Isybot** (2016, France, 2M€), une spin-off du CEA List qui conçoit des robots collaboratifs pour l'industrie, notamment le SYB 3 qui a une portée de 1,6 m avec son bras articulé et une charge utile de 10 Kg.

Marketing et vente

Le marketing et la vente, surtout en ligne, sont l'un des marchés les plus florissants des applications de l'IA. De nombreuses startups fleurissent dans le secteur du marketing et de la vente.

Toutes ne font pas appel à de l'IA même elles sont nombreuses à s'en vanter [995]! Tout le cycle marketing et commercial est couvert par l'IA dans la pratique [996]!

Dans le marketing amont et le planning, l'IA aide à segmenter ses clients, à comprendre leur besoin, à définir des marchés et clients cibles et à interagir directement avec eux. Le profiling d'utilisateurs dans les réseaux sociaux permet de faire du micro-ciblage d'offres.

L'étape de développement de la notoriété tire parti de solutions qui aident à optimiser le plan média et le reach de ses campagnes. Les chatbots interviennent aussi bien en avant-vente qu'en après-vente et leur offre est abondante comme nous l'avons vu précédemment.

Quel est l'impact de ces outils sur les métiers du marketing ? Dans la majorité des cas, ils constituent une boîte à outils étendue pour les marketeurs comme l'ont été de nombreux outils de productivité depuis l'invention de la bureautique et du tableur[997].

---

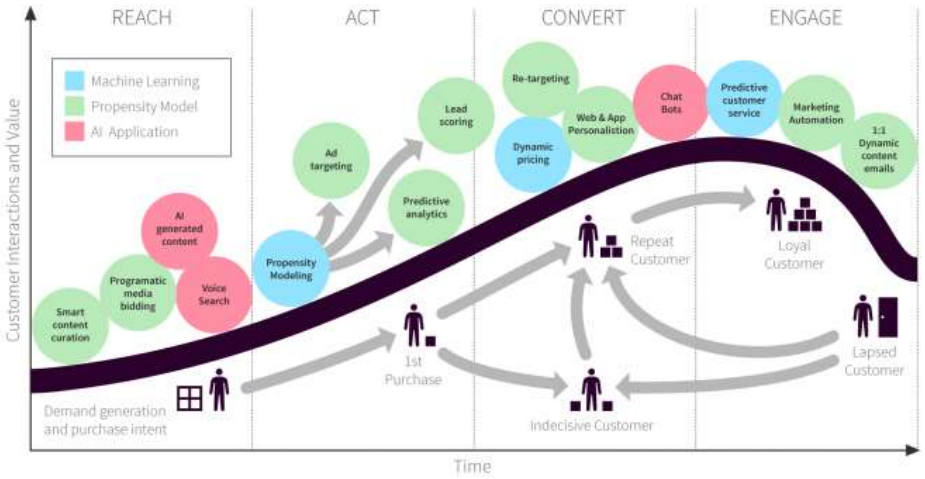
[995] L'excellent Panorama des solutions d'intelligence artificielle pour le marketing publié par Fred Cavazza en octobre 2017 permet d'y voir un peu plus clair (*ci-dessous* à droite) et reste d'actualité en 2019.

[996] Voir Artificial Intelligence for marketing, 2017 (361 pages), un ouvrage librement téléchargeable qui contient un excellent panorama des applications de l'IA dans le marketing.

L'impact peut être plus important dans le cas des chatbots car ils peuvent entraîner une substitution partielle des tâches vis-à-vis d'agents de centres d'appels, surtout pour les appels entrants et le support technique en appels entrants. Mais ils ne doivent pas se substituer entièrement aux interactions humaines qui sont toujours les préférées d'une majorité de clients[998].

planning	notoriété	considération	évaluation	achat	recommandation
segmentation mix marketing pricing	chatbot plan média analyse d'image SEO	chatbot optimisation site web	chatbot	chatbot recommandation upsell / cross-sell	analyse sentiments identification influenceurs

l'IA intervient dans tout le cycle de vente !



Planification

Une bonne part des startups de ce secteur proposent des solutions généralistes intégrant plusieurs outils et accédant à des sources de données externes. Elles servent à segmenter son audience, définir les bons messages et les bons canaux de communication pour les toucher[999].

Albert de la startup **AdGorithms** (2010, IPO en 2015) intègre des outils de segmentation d'audience, d'achats médias, d'optimisation plan média cross-channel, de tests et optimisation et d'analytics. C'est une grosse boîte noire exploitant de nombreuses sources de données. Albert s'appuie aussi sur des briques logicielles d'IBM Watson.

Lucy d'**Equals3** (2015, USA, $7M) s'appuie sur IBM Watson pour segmenter les clients, définir ses messages et optimiser son média planning (vidéo). **Optimove** (2009, USA, $20M) aide aussi les équipes marketing à bien connaître leurs clients et à les segmenter et, dans sa boîte à outils, propose un moyen d'optimiser son mix marketing.

Geotrend (2018, France) s'appuie sur l'IA et en particulier sur le traitement du langage pour automatiser l'intelligence économique à partir des ressources ouvertes du web. Cela permet notamment de réaliser automatiquement une cartographie des relations entre acteurs d'un écosystème (vidéo).

L'optimisation des messages et contenus est aussi le domaine des startups américaines assez bien financées comme **Captora** (2012, USA, $27,3M) et **Persado** (2012, **USA,** $66M).

---

[997] Voir Don't Turn Your Marketing Function Over to AI Just Yet par Kristen Senz, novembre 2019 qui recommande de ne pas croire que les entreprises vont pouvoir remplacer leurs marketeurs par des IA. Comme dans plein de situations, il est recommandé de coupler les IA avec des humains. A bons entendeurs ! Et aussi Council Post: AI Still Thinks More Than It Knows: Three Marketing Missteps To Avoid par James O'Connor, avril 2020, qui est de la même veine.

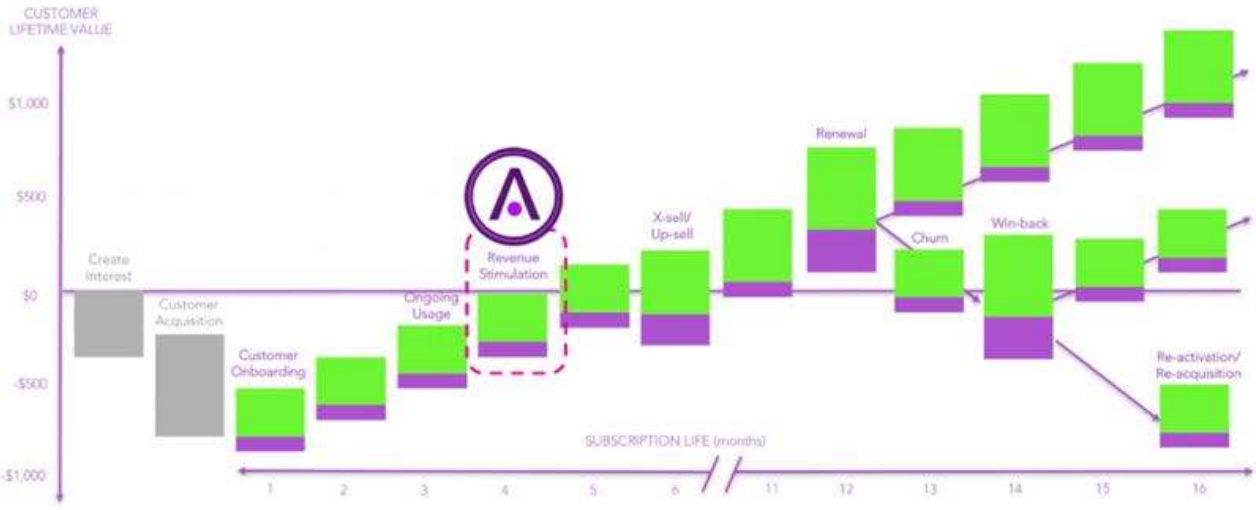
[998] Voir The AI Paradox: Why More Automation Means We Need More Humanity par Michael Brenner, octobre 2019.

[999] Voir AI Makes A Splash In Advertising par Ron Schmelzer, juin 2020, qui évoque l'usage de l'IA pour le ciblage publicitaire.

La planification des messages et des médias s'appuie sur la gestion et l'analyse des données issues des médias sociaux comme avec **Cortex** (2014, USA, $500K) qui prédit la réaction des Internautes aux contenus et **SimpleReach** (2010, USA, $24,2M).

Amplero (2016, USA, $25,6M) aide à ajuster son mix marketing pour générer la plus grande "customer lifetime value" (valeur générée par le client dans la durée). C'est l'outil rêvé pour optimiser ses plans marketing cross-médias. La startup présente quelques études de cas d'optimisations breadth vs depth dans la portée d'actions marketing, cross-sell vs upsell, bref pour gérer les choix cornéliens classiques du marketing.

Dynamic Yield (2011, Israël/USA, $105,3M) a l'air d'être positionné d'une manière similaire pour optimiser un mix média. Elle a été acquise par McDonalds en mars 2019.



Création

D'autres startups sont spécialisées dans la création de contenus avec une assistance de l'IA, souvent très peu décrite dans sa nature exacte.

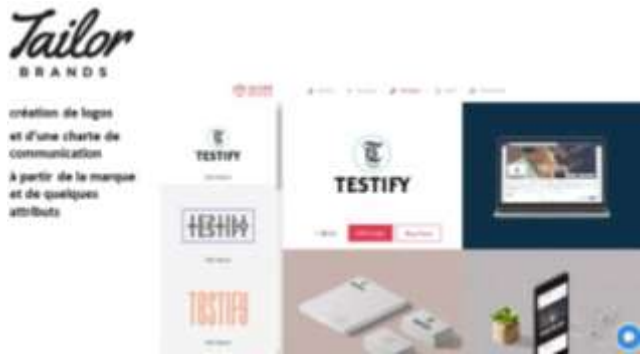
Le deep learning est utilisé dans la création graphique ou la création de sites web. C'est ce que réalise le site de création de sites web **Wix** (2006, Israël, $58,5M, cotée au Nasdaq, vidéo). Il s'est entraîné en analysant le style de nombreux sites web et leur performance (SEO, trafic) pour en déduire empiriquement les paramètres d'un bon site, le tout, en fonction de plusieurs milliers de professions[1000] (vidéo).

L'outil de création de sites web va également exploiter l'ensemble des textes et visuels disponibles sur la société pour bâtir une maquette de site. La création du site est guidée par l'utilisateur qui va tout de même choisir des attributs de sa marque pour influencer le design généré et ses couleurs dominantes. Le danger de ces outils ? Ce n'est pas une surprise : il risque de faire converger les sites créés vers un design voisin.

Une approche voisine est proposée pour l'instant de manière expérimentale par **Adobe**[1001]. Le tout s'intègre dans l'Adome Experience Manager CMS (outil de création de site) et exploite Adobe Sensei, la plateforme logicielle qui intègre les briques d'AI de l'éditeur. Cette IA est focalisée sur l'ajustement des contenus graphiques, pour mieux cadrer les illustrations.

La création à base d'IA porte aussi sur celle de logos. **Looka** (2016, USA, $5,5M, anciennement Logojoy) permet d'en créer avec une méthode un peu bourrine : l'utilisateur sélectionne d'abord cinq styles qui lui plaisent, indique le nom de sa société puis son slogan et enfin, choisi une couleur et cinq icones qui correspondent à son service. A partir de là, le système génère un logo. Ce n'est pas très créatif puisque le logo associe le nom de la société, son slogan et tout ou partie des icones sélectionnées.

Tailor Brands (2014, Israël/USA, $20,6M) aide aussi à créer son logo et sa charte de communication après avoir fourni quelques attributs de sa marque comme son nom et son secteur d'activité[1002]. Ce n'est pas bien sophistiqué mais ça marche. La dose d'IA dedans ? Pas évidente. La startup vise le marché de volume des PMEs qui ne peuvent pas de payer une agence de création de logo traditionnelle. Le tarif est au mois !



Designhill (2014, Inde) fait la même chose (vidéo) mais peut aussi s'appuyer sur des créatifs humains, là où ils ne sont pas chers, pour des tâches de création plus sophistiquées que la création de logos. Ici, le prix de la création de logos via l'IA est fixe, compris entre $15 et $150. Et on obtient un résultat en une minute.

L'agence de communication **McCann** de Tokyo annonçait en 2016 utiliser une IA baptisée AI-CD β comme directeur de création[1003]. Cette IA était construite dans le cadre du « creative genome project" de l'agence. Elle aurait créé un script de publicité TV pour vendre des dragées à la menthe de la marque Clorets du groupe britannique Cadbury. Au bout du compte, celui-ci était de moins bonne qualité que celui du créatif humain de l'agence. AI CD β est matérialisé par un petit système qui dessine des idéogrammes.

[1000] Voir Can You Teach AI To Design? Wix Thinks So par Diana Budds, juin 2016.

[1001] Voir Adobe Is Building An AI To Automate Web Design. Should You Worry?, mars 2017.

[1002] Voir une description du processus dans Avec l'IA, la génération de logos passe au low-cost de Lélia De Matharel, décembre 2017.

[1003] Voir Human Beats AI CD in McCann Japan's Creative Battle, septembre 2016. Voir cette vidéo qui explique l'histoire.

Il en est allé de même en septembre 2016 pour des publicités réalisées par **Burger King** avec une IA générant automatiquement le scénario (vidéo)[1004]. Visiblement, l'IA en question analysait une base de 10 ans publicités TV bien labellisée avec les études d'impact pour identifier leurs éléments d'efficacité. Cette base était croisée avec les éléments du brief de la publicité comme l'audience visée, l'objectif de la campagne et le « claim » de la campagne. L'IA faisait des recommandations sur les éléments de l'annonce comme le contexte, la tonalité musicale ou le type de célébrité à y intégrer. Le reste était réalisé par des collaborateurs de l'agence. Bref, en guise d'IA, nous avions une application de machine learning multicritère.

Human Beats AI CD in McCann Japan's Creative Battle

By Erik Oster on Sep. 1, 2016 - 10:27 AM

After introducing its A.I. CD early this spring, McCann Japan decided to bit the AI-CD β robotic creative director against its human counterpart, namely creative director **Mitsuru Kuramoto**, in a creative battle. Both were given the task of creating a spot for Mondelez Japan brand Clorets Mint Tab, communicating the brand message of "instant, long-lasting refreshment that lasts for 10 minutes" and then turning to a nationwide poll to declare the winner.

En octobre 2017, l'agence Belge **DDB** faisait jouer le rôle de juré à une IA développée avec Microsoft pour sélectionner la meilleure publicité dans la dixième édition de leurs MIXX Awards. Le Video Indexer Service de Microsoft servait à analyser dix ans d'archives publicitaires taggées pour prédire lesquelles seraient en haut du panier[1005].

Kuaizi Technology (2013, Chine, $1,5M) propose enfin la création de contenus basée sur l'analyse de la performance de contenus anciens. La société était présente sur le stand de LVMH de Viva Technology en mai 2018.

C'est un partenaire de Facebook. Leur « Kuaizi PCP Programmatic Creative Platform » permet notamment de faire de l'A/B testing de créations sur Facebook. Et leur « SEM Smart Matching System of Landing Page » sert à optimiser la page d'accueil d'un site web. Il utilise probablement également du machine learning, même si ce n'est pas précisé.

Web et analytics

Nous avons aussi des solutions d'optimisation de sites web comme **Webpage.ly** (2015, Canada) qui est focalisé sur le référencement naturel (SEO) et fonctionne en mode cloud. **Tilofy** (2013, USA, $1M) est de son côté une solution solution de prévision des tendances dans leprêt à porter et les usages, dont les méthodes et techniques ne sont pas précisées mais qui doit faire appel à des techniques d'analyse du langage (NLP).

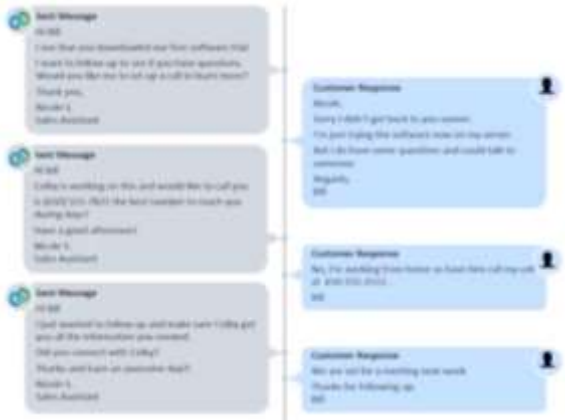
[1004] Voir Burger King's AI-written ads are beautiful disasters, septembre 2018.

[1005] Voir Meet Pearl – an AI judge who is already dishing out awards, octobre 2017.

Network Insights (2006, USA, $77,2M, acquis par **American Family Insurrance** fin 2017) propose Audience.ai, un outil d'analytics qui exploite les traces des utilisateurs dans les réseaux sociaux pour définir des profils de prospects et clients ultra-précis et des messages associés ultra-ciblés. Ce n'est pas sans rappeler les méthodes de feu Cambridge Analytica qui avait joué un rôle pour cibler les messages pro-Trump – ou décourangeant les minorités de voter - dans les *swing states* pendant la présidentielle américaine de 2016 et défrayé ensuite la chronique au printemps 2018 lorsque les méthodes associées ont été dévoilées au grand public, mettant en cause au passage Facebook.

Cette approche marketing consistant à faire du micro-ciblage multi-factoriel sur des clients avant de lancer ses campagnes marketing est aussi proposée par une startup de Palo Alto, **Mariana** (2014, USA, $4M).

Meltwater (2001, USA, $60M en dette) propose des solutions en cloud de veille et d'analyse de l'information sur les médias en ligne et sociaux. Elle croit rapidement par acquisitions, avec quatre acquisitions en 2017 dont celle de **Wrapidity** (2015, UK, £50K), issue d'un projet de l'Université d'Oxford. Meltwater couvre la veille stratégique, la pige média en ligne, le ciblage de journalistes, l'e-réputation, l'analyse des réseaux sociaux et de sentiments sur les marques et la mesure de performance des campagnes marketing en ligne. Le tout est présenté dans un tableau de bord (*exemple ci-dessous à gauche*).



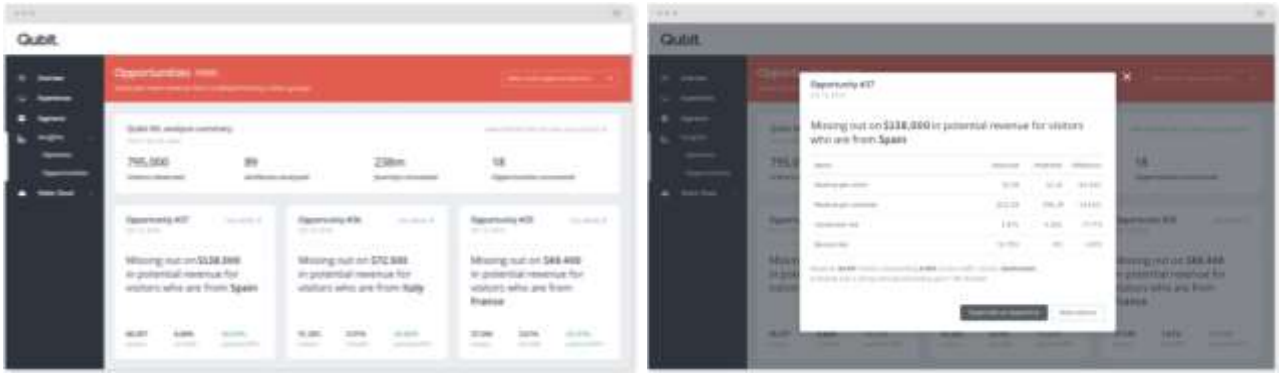
Conversica (2007, USA, $72M) est un outil d'automatisation de la communication par mail à des prospects (*exemple ci-dessus à droite*). Comme d'habitude, les techniques d'IA ne sont pas précisées mais relèvent certainement de combinaison d'outils de NLP (traitement du langage).

L'outil s'interface comme il se doit avec ceux de SalesForce et met le client en relation avec un véritable conseiller commercial si nécessaire. C'est une sorte de Julie Desk appliquée aux premiers traitements d'une demande d'un prospect. Une explication du processus dans cette vidéo.

Spella (2014, France/Belgique) collecte et structure les données issues des commentaires des internautes pour identifier les signaux forts et faibles avec un bon niveau de granularité.

Support aux ventes

En complément du précédent, **People.ai** (2016, USA, $37M) fournit de son côté des outils d'aide et de coaching de conseillers commerciaux pour les former à closer les deals, par analyse de patterns d'appels antérieurs, et aussi pour les éviter de passer trop de temps avec les clients impossibles à closer. Leur outil analyse plus de 90 sources d'information dont les outils de communication de l'entreprise (mail, vidéoconférence, etc) pour alimenter le CRM. Le machine learning analyse les textes, les émotions et les activités pour identifier des points à régler dans le cycle de vente afin de l'accélerer. La startup propose aussi des vidéos de formation pour conseillers commerciaux.



Qubit[1006] (2010, UK, $74M) a développé un moteur de détection automatique des opportunités de revenu à base de machine learning, exploité principalement par les sites de vente en ligne. C'est en fait un outil de segmentation automatique de clients pour identifier ceux qui sont les plus prometteurs. Le système permet aussi de piloter des campagnes en ligne d'A/B Testing d'offres commerciales ciblées.

LeadGenius (2011, USA, $19M) propose une solution intégrée de génération, de qualification et de conversion de leads en mode b2b. Elle s'appuie sur le scrapping de sources Internet diverses (LinkedIn qui contient plus de 7,5 millions d'entreprises aux USA, Yelp, AngelList et Crunchbase pour les startups, déclarations de résultats auprès de la SEC, base d'organisations non-profit de l'IRS). Ils utilisent des techniques de segmentation par machine learning pour dédoublonner la base. En gros, c'est l'équivalent d'une base Compass générée par scrapping. Leur IA permet aussi d'analyser le comportement des clients pour identifier leurs besoins via un réseau de neurones entraîné avec des données comportementales de milliers de clients. Enfin, ils engagent ensuite la conversation avec eux via des mails personnalisés et en analysant les réponses des clients pour gérer la suite des événements[1007].

Alphalyr (2014, France) a créé une solution de reporting commercial qui s'appuie sur de la business intelligence et de l'IA. Elle génère un email de synthèse quotidien avec les indicateurs clés. Ils utilisent des réseaux de neurones pour l'apprentissage et la détection de signaux faibles. C'est certes assez vague, mais correspond bien aux techniques du machine learning non supervisé.

Showpad (2011, Belgique, $159,5M) a développé un outil de coaching et d'outillage des commerciaux de sociétés plutôt orientées b2b.

[1006] Il est dommage que cette société utilise une dénomination liée à l'informatique quantique, qu'elle n'utilise pas du tout. Quand on cherche des vidéos les concernant, on tombe bien évidemment en premier sur des explications sur les qubits de l'informatique quantique et pas sur leurs propres vidéos. C'est bien ballot !

[1007] Voir la vidéo Data Driven Sales: Building AI that searches, learns and sells de Anand Kulkarni, leur Chief Scientist, septembre 2015.

IBM va jusqu'à faire appel au machine learning pour sélectionner ses partenaires pour leur envoyer des leads commerciaux, avec l'outil **Score** (Smarter Cognitive Opportunity Recommendation Engine)[1008]. Il s'appuie sur l'historique des leads passés. Cela doit renforcer les meilleurs partenaires et éliminer les autres progressivement. Le système est censé avoir apporté $100M de CA supplémentaire à IBM et accéléré la vitesse de traitement de ses leads.

Le spécialiste de l'équipement des centres d'appels **Genesys** est censé utiliser IBM Watson pour améliorer ses services en analysant le flot de données généré par les appels clients. C'est aussi ce que fait **Daisee** (2017, Australie, $6,7M).

Histoire de boucler la boucle, reste à vérifier, par analyse des sentiments, que cette montée en charge de l'utilisation de l'IA dans la relation clients n'aboutisse pas à un rejet de ces mêmes clients excédés par les robots, comme la sélection directe à l'arrivée (SDA) dans les centres d'appels !

Afiniti (2006, USA, $197M) propose une IA à base de machine learning pour apairer clients et agents de centres d'appels. Leur méthode de machine learning non documentée exploite les données disponibles sur les clients. Ils ont déposé 150 brevets pour protéger ces méthodes.

Q°emotion (2015, France, 814K€) se positionne comme une plateforme « d'insights émotionnels » obtenus à partir de l'analyse sémantique d'avis clients publiés sous forme de commentaires dans les réseaux sociaux, dans les emails ou les verbatims d'enquêtes clients. L'outil est censé aider à contextualiser les émotions pour améliorer les parcours client. La solution est fournie en SaaS. Ils ont quelques clients en France dans la banque et le retail mais leur solution est déjà multilingue.

DeepReach (2017, France, $9,7M) propose une solution d'optimisation de la communication locale des marques pour les agences et les annonceurs lancée par Jean-Pierre Remy, l'ancien PDG des Pages Jaunes. La solution exploitant le machine learning permet d'optimiser les campagnes locales, réduisant significativement leur temps de préparation.

Personnalisation

La personnalisation de l'expérience client est un objectif de nombreuses marques, revendeurs et annonceurs et les startups ne manquent pas d'imagination pour mettre en scène l'IA pour proposer des solutions allant dans ce sens. C'est le vaste business de la personnalisation et de la recommandation. Mais il y a du boulot car il est très difficile de disposer des informations relatives à chaque client ou prospect pour leur faire des recommandations pertinentes, surtout en termes de timing.

On vit au quotidien ces erreurs de ciblage avec d'innombrables scénarios de ratés. Nous sommes bombardés de reciblage publicitaire dès que l'on a cherché un produit d'un type donné. Ce harcèlement est risible.

Il intervient bien évidemment même lorsque l'on a acquis le produit en question, ce que les autres sites et ad exchange ne savent pas forcément. C'est le cas pour les recommandations dans le voyage.

Vous planifiez vos vacances pour une destination donnée, vous y allez, et vous serez ensuite exposé à de la publicité sur cette destination pendant des années, même si vous changez à chaque fois de destination. Bref, le marketing prédictif ne sait pas gérer vos envies imprédictibles ni même vous faire de recommandation pertinente pour une autre destination !

Les méthodes les plus efficaces consistent à s'intégrer là où les clients expriment des besoins comme dans les moteurs de recherche, qu'il s'agisse de Google Search ou Bing, ou dans des sites de vente en ligne généralistes comme Amazon.

[1008] Voir IBM s'appuie sur Watson pour gérer ses partenariats par Bastien Lion, mai 2019.

L'IA permet-elle d'envoyer un message personnalisé ? En théorie oui. Mais il faudrait qu'elle dispose de bien plus d'informations qu'elle n'en collecte aujourd'hui, sauf chez Google qui sait le plus de choses sur notre vie d'internaute.

Il faudrait aussi assembler des outils bien plus sophistiqués que ceux qui existent aujourd'hui. Comme pour notamment analyser les sentiments, le vocabulaire employé par le client lorsqu'il s'exprime dans les réseaux sociaux et s'y adapter de manière… manipulative par psycho-engineering. Les solutions en reviennent souvent aux techniques de micro-segmentation plutôt que de véritable personnalisation.

Les outils de recommandation de produits dans les sites de vente en ligne s'appuient sur du machine learning. Les catalogues de produits sont valorisés avec des systèmes de reconnaissances d'images similaires. Les sites web tirent parti d'outils d'optimisation du parcours utilisateur. Enfin, de nombreux outils automatisent ou accompagnent le rôle des conseillers commerciaux dans les centres d'appels entrants et sortants.

Emarsys (2000, Autriche, $55,3M) propose des outils de personnalisation tout azimut de la relation client multicanal, mis notamment en œuvre dans l'envoi d'emails ciblés, avec de l'A/B Testing s'appliquant notamment à l'heure idéale d'envoi d'emails promotionnels. Leurs outils permettent aussi de personnaliser la navigation dans le site web de la marque et notamment pour ce qui est des scénarios de sortie.

SouthPigalle (2015, France, $1,8M) vise aussi la personnalisation de l'expérience client. Ils ciblent notamment les entreprises du secteur du luxe. Ils proposent une boîte à outils intégrant la création de chatbots, des outils de segmentation de bases clients et de marketing prédictif pour automatiser la création de parcours client personnalisés.

NSP (2000, France) est une solution d'automatisation du marketing en mode cloud. La startup est partenaire d'Inria autour du machine learning appliqué au monde du tourisme. On est dans le cas d'une entreprise plutôt traditionnelle qui cherche à moderniser ses outils.

Vestorly (2012, USA, $14,6M) propose un moteur de curation de contenus issus du web exploitant des techniques de machine learning pour la recommandation de contenus personnalisés. Il est notamment associé à la plateforme d'emailing de l'éditeur. Leur solution est intégrée à HootSuite depuis 2019 et notamment dans la plateforme de publication Hootsuite Amplify.

L'écosystème français a toujours été prolixe en startups b2b et b2c dans le secteur du ecommerce et du marketing. Il est donc normal d'y retrouver quelques startups intégrant de l'IA.

AntVoice (2011, France, 4,1M€) propose une solution de recommandation prédictive pour les sites de ecommerce qui s'appuie sur du machine learning. La solution analyse la pondération de la relation entre Internautes et produits et s'appuie sur la théorie des graphes. Ils ont pivoté pour cibler le marché de la publicité en ligne.

Datapred (2014, France) propose également une solution d'analyse prédictive basée sur du machine learning. La société cible divers marchés professionnels dont celui de la distribution, en plus de la finance, de la logistique et de la santé. Elle permet par exemple de simuler des hypothèses marketing et leur impact sur une chaîne logistique de distribution en tenant compte d'un grand nombre de paramètres. Comme c'est souvent le cas, le lancement d'un projet requiert une bonne part de service et de personnalisation avant sa mise en oeuvre opérationnelle.

DataPublica / C-Radar (2011, France) est une société qui propose une solution en cloud de marketing prédictif B2B permettant de cibler les bons prospects. Elle s'appuie sur l'exploitation des données administratives et financières des entreprises issues de sources publiques, des sites web associés, des réseaux sociaux et des mentions dans les médias.

Ces données permettent alors de segmenter automatiquement les clients, de priorisation de ces segments, le tout s'appuyant sur un apprentissage supervisé. L'approche permet par exemple de segmenter les startups d'un secteur d'activité donné (Medtech, Fintech). La société est une autre participation d'IT Translation. Elle a été acquise par l'éditeur de logiciels **Sidetrade** (2000, France) en juin 2017. Ce dernier propose maintenant outils prédictifs et décisionnels, exploitant du deep learning et du traitement du langage (NLP) pour accompagner les entreprises sur l'ensemble du cycle de vente, comprenant l'identification de marché cible, les opportunités d'upselling et la gestion prédictive des délais de paiement[1009].

D'autres startups se positionnent sur ce créneau comme **Compellia** (2015, France), qui analyse des sources de données ouvertes et identifie des événements clés de la vie des entreprises pour créer des listes de prospects qualifiés, sachant que le processus est spécifique à chaque marché.

Il y a aussi **TinyClues** (2010, France, $25,4M), une startup plus établie qui utilise des solutions de machine learning pour identifier les produits que les clients de sites de vente en ligne sont les plus susceptibles d'acheter, histoire d'optimiser les campagnes marketing ciblées au niveau du ciblage comme des messages et des offres. A Camaieu comme client.

Dictanova (2011, France, $1,3M) est une société nantaise à l'origine d'une solution d'analyse textuelle des feedbacks clients dans les réseaux sociaux ou sites de vente en ligne, en liaison avec les outils de CRM pour optimiser la relation client. Les techniques utilisées comprennent l'analyse sémantique de textes et la classification automatique. La solution est fournie en cloud. C'est une autre participation d'IT-Translation.

Frame AI (2016, USA, $10,6M) analyse les conversations clients par exploitation des données des systèmes de gestion de centres d'appel, d'outils comme Slack et Zendesk et de SFA/CRM comme Salesforce. Il permet notamment d'automatiser le processus d'escalade de problèmes clients, de détecter des tendances dans les incidents et de faire du reporting. Le tout exploite surtout de l'extraction de données des textes des conversations.

Cresta (2017, USA, $21M) utilise aussi le traitement du langage dans les centres d'appels mais pour suggérer aux agents des réponses appropriées basées sur celles qui fonctionnent le mieux.

Nalia (2020, France) est une startup qui veut aider les entreprises à détecter le churn de leurs abonnés ainsi que le potentiel d'upsell en exploitant les données captées par les systèmes de relation client et de centres d'appels. Elle produit de nombreux tableaux de bord permettant d'avoir une vue à 360° d'un portefeuille de clients.

Modizy (2012, France, $275K) propose un assistant d'achats dans la mode basé sur un algorithme d'intelligence artificielle. Modizy propose aussi une place de marché reliant consommateurs et marques.

Enfin, la personnalisation touche aussi les emailings. L'IA peut servir à envoyer les bons contenus aux bonnes personnes et au bon moment[1010].

Evénementiel

L'organisation d'événements comme des conférences et des salons peut aussi faire appel à des outils exploitant l'intelligence artificielle. C'est un secteur encore naissant mais porteur d'innovations.

[1009] Voir leurs vidéos d'études de cas. Leurs solutions sont notamment déployées à la Direction Financière de Manpower en France pour le recouvrement de factures.

[1010] Voir Next-Gen Email: Can AI deliver higher engagement? par Brian Kardon, mai 2019.

Il peut s'appuyer sur la capacité à capter des éléments de comportement de l'audience ou des visiteurs, pour les analyser et orienter ces derniers.

Nous avons déjà vu le cas de **Datakalab** qui analyse les émotions visibles de l'audience d'un événement, permettant d'identifier les messages et intervenants les plus impactants de manière chiffrée (on s'en rend compte souvent sans outil…).

Il y a aussi **Bziit** (France), une startup bordelaise qui veut associer l'IA et l'événementiel. Leur plateforme logicielle sert à traiter et classifier les données événementielles collectées pour détecter diverses anomalies sur le visitorat en fonction des investissements dans la promotion et le ciblage, analyser l'audience et ses réactions, son profiling avant et après l'événement, le tout avec génération de recommandations pour optimiser le rendu d'un événement.

Les données exploitées sont celles des flux physiques (visitorat par salon ou stand, durée de visites, timing), les flux dans les réseaux sociaux (discussions, hashtags, tonalité, profil des influenceurs) et les flux CRM (contacts générés, leads, etc). Ils visent les marchés du tourisme, de la distribution, des collectivités locales et de l'événementiel.

Il faudrait y ajouter une catégorie déjà bien couverte, celle des robots d'information pour les lieux publics, qui font aussi partie d'une stratégie marketing.

Relations Publiques

Les agences et équipes en charge des relations presse et influenceurs peuvent aussi faire appel à quelques outils mâtinés de briques d'IA.

L'un d'entre eux est proposé par **AlgoLinked** (2014, France) qui fait de la qualification des fichiers de journalistes et influenceurs en croisant le contenu d'un communiqué de presse avec le passé rédactionnel ouvert des journalistes référencés. La startup propose aussi une offre aux journalistes pour les aider à trouver des experts.

Le traitement du langage pourrait être mis à contribution pour rendre la communication vers les médias plus efficace. On pourrait imaginer voir apparaître un détecteur de pourcentage de bullshit dans un communiqué de presse (pour les lecteurs) et qui prodiguerait des conseils pour en réduire la proportion (pour les producteurs), comment avec des preuves, des termes plus précis ou des données chiffrées. On peut rêver un peu ? Ce genre d'outil est tout à fait possible à réaliser, mais le marché des agences de relations publiques est peut-être trop étroit pour justifier la création d'une startup. Resterait à le proposer pour les créations marketing, ce qui donnerait un marché cible un peu plus grand et plus solvable.

Ressources humaines

Peut-on injecter de l'intelligence artificielle dans les ressources humaines ? Il semble que oui, tout du moins, essentiellement dans les processus de recrutement.

C'est encore en observant les créations de startups que l'on peut se faire une idée des grandes tendances du domaine, surtout aux USA, où les entreprises n'ont peur de rien et ne se soucient pas forcément d'éthique ou de valeurs humaines, malheureusement.

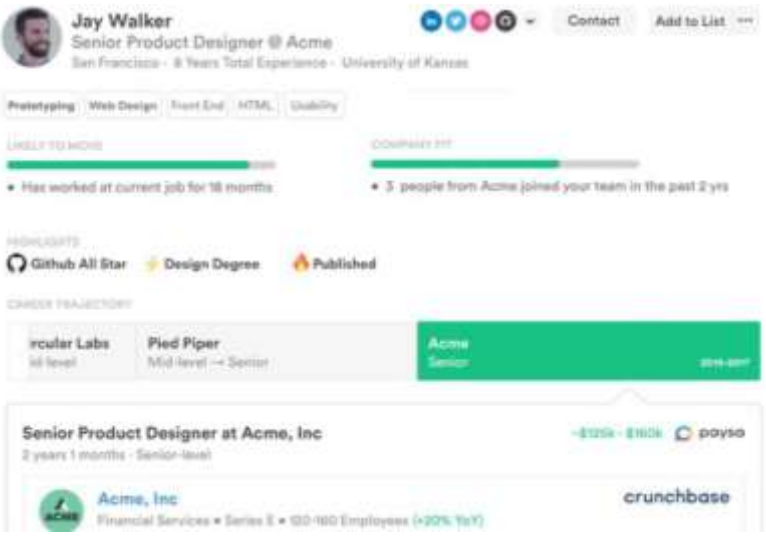
En plus du recrutement, l'IA dans la RH peut aussi servir à gérer les talents internes de l'entreprise, à les affecter à des missions et à gérer leur mobilité interne.

A y regarder de près, il vaudrait mieux que ces différents outils facilitent le travail des RH et des recruteurs plutôt qu'ils ne les remplacent car sinon, à ce train là, les IA auront un rôle nivelant dans le recrutement et tous les originaux se feront éjecter sans compter d'éventuelles discriminations qui pourraient provenir des biais volontaires ou involontaires des données d'entraînement de ces IA[1011]. Est-ce que l'IA rend le recrutement plus humain[1012] ? Pas vraiment ! Mais comme partout ailleurs, l'irruption de l'IA dans la RH semble inéluctable même si son adoption est encore relativement lente dans les entreprises[1013].

Recrutements

Nombre de startups proposent des solutions qui permettent d'améliorer la qualité et la productivité des recrutements[1014]. Cela commence avec des outils d'aide à la rédaction d'annonces d'emplois efficaces et d'analyse des réponses des candidats comme chez **Textio** (2014, USA, $29,5M). On est ici dans le domaine du traitement du langage (NLP).

Suivent des outils d'analyse prédictive pour identifier des talents à chasser avec **Entelo** (2011, USA, $41M, *ci-contre*) et **Gild** (2011, USA, $26M, acquis par Citadel en 2016). Ce genre d'outil s'appuie sur des techniques de prévision exploitant du machine learning. Entelo est doté d'un moteur de recherche qui scrute les profils d'individus sur Internet pour les exploiter, à partir de 70 critères comme l'état de leur employeur (acquisition, IPO, évolution du cours de bourse, analyse de sentiment).



La partie IA de ce genre de solution n'est pas visible par les candidats sollicités. Ce sont des outils d'*empowerment* des recruteurs. La startup française **Clustree** (2013, France, $11,6M), lancée par Bénédicte de Raphélis Soissan, utilise aussi l'IA pour rapprocher l'offre et la demande.

[1011] Voir l'ouvrage collectif L'IA au service des RH – pour une expérience collaborateur augmentée, septembre 2020 (258 pages), avec les contributions de Michel Barabel, Charles-Henri Besseyre des Horts et Timothée Ferras, Thierry Bonetto, Charlotte du Payrat, Magali Mounier-Poulat et Gaëlle Bassuel. Il décrit les usages de l'IA dans la RH avec les précautions éthiques qui doivent être associées pour ne pas enlever le H aux RH et éviter que les IA RH véhiculent et amplifient les biais de la société.

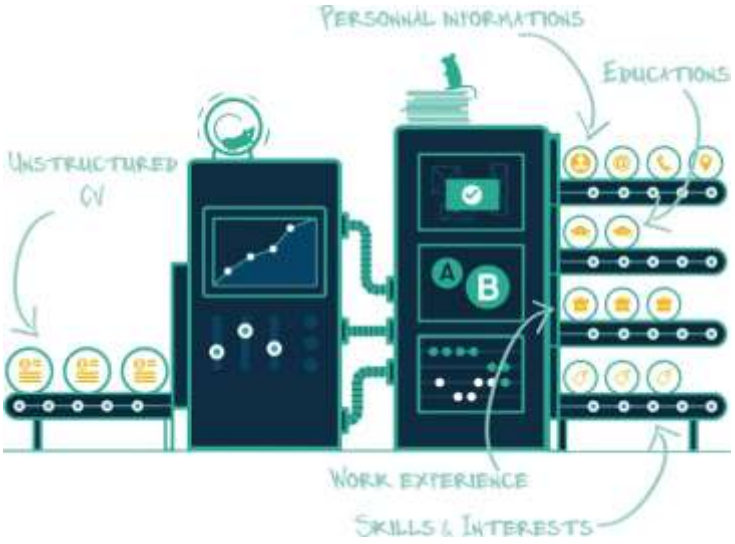
[1012] C'est la thèse défendue dans How AI Makes Recruiting More Human de Steven Jiang en septembre 2018.

[1013] Voir AI in HR seen as inevitable -- and unstoppable par Patrick Thibodeau, janvier 2019.

[1014] Voir L'intelligence artificielle peut-elle aider à recruter de pépites ? par Pierre Maurin, mars 2019 ainsi que 3 Ways AI Is Changing The Game For Recruiters And Talent Managers par Rebecca Skilbeck, décembre 2019.

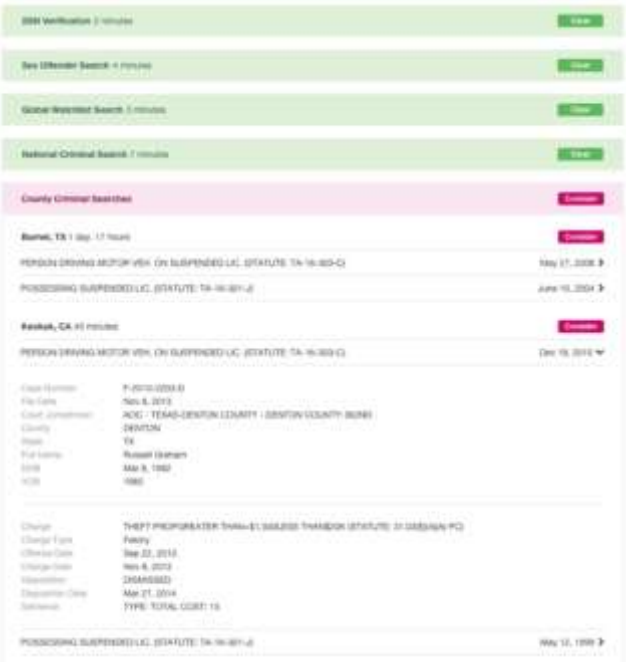
Riminder (2016, France, $2,3M) est une autre startup qui propose de filtrer les CVs avec du deep learning en récupérant les informations structurées et non non structurées comprises dans les CV des candidats et dans les ressources d'Internet. Cela leur permet de faire des prévisions sur l'adéquation des CVs aux postes ouverts. Ils utilisent de l'analyse sémantique pour extraire les informations textuelles des CV et de la reconnaissance d'entités (noms d'employeurs, email, téléphone, postes occupés, formation et diplômes). Leur système est entraîné pour reconnaître et rapprocher une grande variété de fonctions. Ils utilisent ensuite du machine learning pour faire de la segmentation automatique pour identifier les meilleurs profils.

Attention cependant aux petits malins qui commencent à connaître les méthodes permettant de bien remplir un CV de mots clés pour que les IA les retiennent [1015]! Il faut aussi se méfier des biais stéréotypant des IA, aussi bien intentionnées et entraînées soient-elles [1016]. Les propositions de candidats sont alors expliquées selon une grille de critères (expérience, soft skills, motivation, formation, etc). Les fonctions de Riminder sont aussi disponibles sous forme d'APIs pour les éditeurs de logiciels de RH. Les données d'entraînement sont issues de sources mondiales pour éviter les biais culturels locaux.



Paradox.ai (2016, USA, $13,3M) a créé un chatbot de recrutement gérant un dialogue qualifiant avec les candidats. C'est adapté aux métiers d'exécution comme dans le retail et la restauration.

Checkr (2014, USA, $149M) est une startup qui réalise des vérifications de CV et analyses de réputation de candidats. Elle aurait déjà plus de 10 000 clients avec son offre en cloud dont Uber. Le principe consiste à scanner toutes les sources publiques disponibles pour identifier les incohérences, bizarreries ou plus simplement, un casier judiciaire. Vu le nombre de prévenus passés par la case prison aux USA, c'est encore plus utile là-bas que dans les pays européens (*ci-contre*).



Bayes Impact (2014, France, $120K) est une société originale qui veut utiliser l'IA pour le bien public et se positionne comme une ONG. Son créateur Paul Duan s'est fait connaître en lançant un partenariat avec Pôle Emploi pour faciliter le rapprochement entre l'offre et la demande d'emplois.

[1015] Voir How to AI-proof your resume par Ben Reuveni, mai 2019.

[1016] Voir quelques exemples de biais dans De l'IA chez les RH : vers le recrutement "prédictif" de profils uniformisés ? par Fabien Soyez, février 2019.

L'offre comprend un agent conversationnel qui aide le chercheur d'emploi à se positionner et à trouver un travail correspondant à ses capacités.

Groupe Gorgé (France) est une PME industrielle spécialisée en robotique civile et militaire. En 2020, ils lançaient StedY.io, un outil à base d'IA qui sert à automatiser le recrutement d'ingénieurs, sachant que StedY est lui-même un employeur sur le marché de l'ingénierie.

IBM Watson est utilisable pour analyser votre personnalité à partir de vos écrits, en s'appuyant sur Personality Insights et Tone Analyzer, deux outils d'analyse de vos écrits[1017] qui font partie des différentes API de Watson. La solution permet en tout cas de détecter l'humeur de l'auteur, comme **sa tristesse**. Et peut-être d'améliorer les recrutements, tout du moins de candidats qui ont une vie publique sur Internet. L'analyse de personnalité peut aussi exploiter les vidéos d'interviews, même si le candidat parle à une machine, comme un jeu de serious gaming.

C'est ce que propose l'Américain **HireView** (2004, USA, $93M). Leur logiciel analyse les exppressions faciales dans les visages et identifie des traits de personnalité[1018]. La solution est déployée chez Unilever aux USA[1019] et au Royaume Uni[1020]. L'efficacité de ce genre de recrutement à la tête du client serait sujette à caution[1021].

L'histoire pourrait se compliquer si les recruteurs se mettaient à utiliser des variations de ces systèmes d'analyse du visage comme ce prototype de Stanford qui détermine automatiquement les préférences sexuelles[1022].

Le recrutement est aussi un autre terrain de jeu pour les chatbots comme celui que l'agence française **TheChatbotFactory** a déployé à la BNP.

Aux USA, **Google** est aussi présent sur ce marché avec Google Hire lancé en juillet 2017 et surtout destiné aux PME qui utilisent G Suite, la suite bureautique en ligne de Google. Le processus de recrutement s'intègre donc naturellement dans Gmail et Google Calendar pour dispatcher les entretiens d'embauche vers les personnes disponibles dans l'entreprise. Google Hire permet aussi d'optimiser le texte de ses offres d'emploi, ne serait-ce que pour l'indexation dans Google Search ! Google Hire gère enfin un vivier de candidats pour les faire ressortir du lot lors de nouvelles ouvertures de postes. A vue de nez, c'est plus une application de travail collaboratif qu'une IA mais

---

[1017] Voir IBM Watson Developer Cloud, Personality Insights et IBM Watson Developer Cloud, Tone Analyzer.

[1018] Il existe divers moyens d'analyse. Des chercheurs arrivent à le faire juste en analysant le mouvement des yeux ! Voir Eyes tell all for artificial intelligence - even your personality, new study finds, juillet 2018.

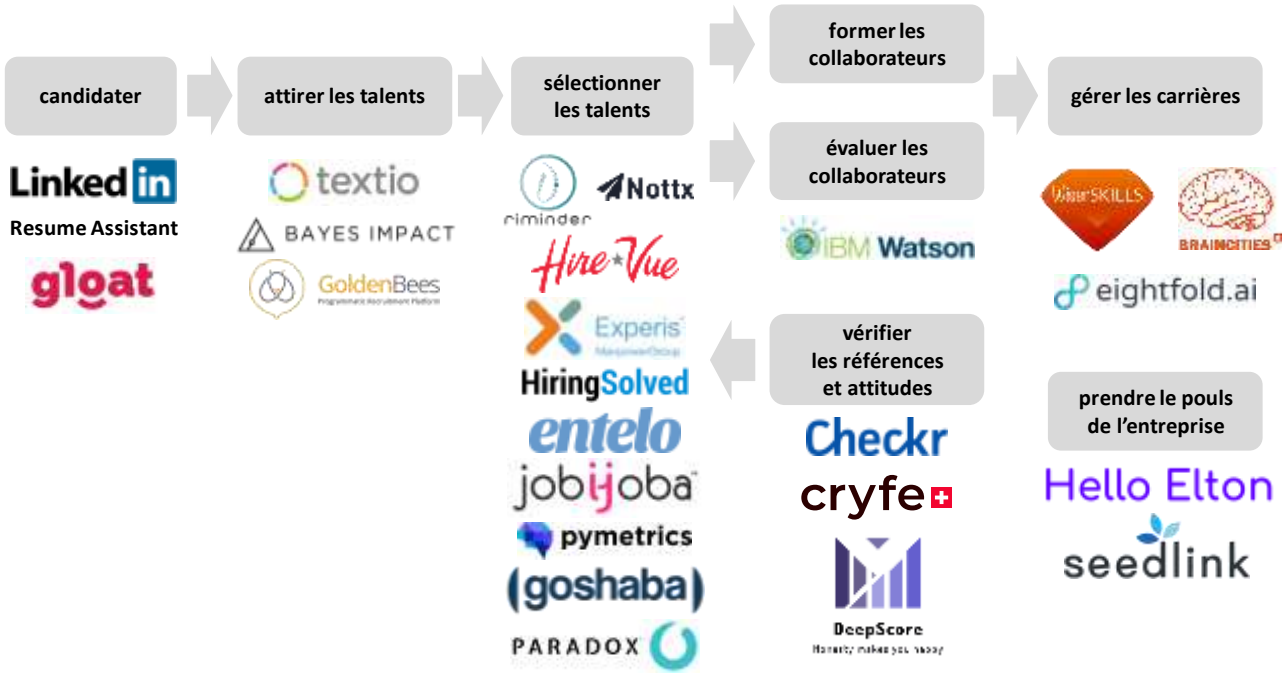
[1019] Voir l'étude de cas documentée par HireVue : https://www.hirevue.com/customers/unilever-finds-top-talent-faster.

[1020] Voir AI used for first time in job interviews in UK to find best applicants par Charles Hymas, septembre 2019.

[1021] Selon Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices par Manish Raghavan et al, juin 2019 (24 pages) selon lequel ces systèmes sont fragilisés par les biais issus des données d'entraînement.

[1022] Voir This AI knows whether you're gay or straight by looking at a single photo, septembre 2017.

Google met en avant le fait qu'elle s'appuie sur de l'IA, ne serait-ce qu'au niveau de ses fonctions d'analyse du langage.



D'un point de vue symétrique, **Microsoft** et sa filiale **LinkedIn** utilise sa base pour conseiller les utilisateurs lors de leur rédaction de CV (vidéo) dans Word, le Resume Assistant. Ce n'est qu'un des outils que l'éditeur pourrait proposer afin de faciliter le recrutement. Et comme ils disposent de la base de LinkedIn, ils peuvent l'exploiter à fond !

Experis IT, une filiale de Manpower, a développé une solution qui permet d'avoir un entretien d'embauche avec un avatar, qui n'est pas sans rappeler les tentatives du genre réalisées par L'Oréal et le Crédit Agricole en 2007 sur Second Life. Leur Digital Room qui est un co-con d'isolation du candidat analyse son visage et sa voix ainsi que ses réponses aux questions posées. Le système va alors générer un rapport au recruteur et au candidat. C'est visiblement adapté au recrutement de fonctions d'exécution. Ca fait envie ! Cette innovation digitale était présentée lors de Viva Technology en mai 2018. L'inconvénient est de nécessiter cet engin, mais ont devrait pouvoir faire la même chose à distance avec un simple laptop.

Goshaba (2014, France) propose une solution faisant la correspondance entre CV de candidats et de postes à pourvoir pour les recruteurs. Ils proposent aussi un jeu mobile qui sert à évaluer les compétences sociales des candidats. Il y a peut-être de l'IA dedans, mais sans qu'elle soit bien facile à identifier.

Merito (2016, France, 2,2M€) aide à recruter des travailleurs à temps partiel pour la distribution. Le matching exploite les notions de distance ainsi que l'évaluation des précédents managers.

Cryfe (2020, Suisse) est un logiciel d'analyse comportementale destiné au RH et autres professionnels qui pourraient en avoir besoin (psychologues, communication, politiques...). Il exploite le deep learning dans le traitement du langage et l'analyse d'images pour détecter l'authenticité de l'interlocuteur. Une solution équivalente est proposée par **DeepScore** (Japon).

Golden Bees (2015, France, 1M€) propose la solution Wan2Bee qui permet le ciblage de candidats via de la publicité programmatique. L'outil est destiné aux recrues potentielles qui voient des annonces de postes ouverts à la place des publicités en ligne classiques.

Le service est complété par un outil qui permet de candidater rapidement à plusieurs postes. Le candidat a d'abord créé son profil sur le site de Wan2Bee. Mais pourquoi passer donc par de la publicité au lieu de la consultation d'offres dans le site ? Pour éviter d'avoir à le consulter ?

La grande question qui se pose avec toutes ces IA dédiées au processus de recrutement est les biais qu'elles pourraient véhiculer voire renforcer[1023]. Les données d'entraînement des bases contiennent structurellement des biais de la société. On peut éventuellement les supprimer d'emblée en anonymisant les CV comme le propose **Nottx** (2015, UK) avec sa solution Fairhire, qui s'appuie sur le l'IA qui identifie les parties des CVs qui doivent être supprimées comme l'identité, l'origine sociale et géographique ainsi que les formations suivies. L'autre solution consisterait à biaiser dans un sens positif et contrôlé les données d'entraînement, par exemple, en renforçant le poids statistiques des populations trop peu présentes dans les métiers. C'est ce que semble proposer **pymetrics** (2011, USA, $56M) qui vise explicitement à augmenter la diversité dans le recrutement dans lesentreprises.

Enfin, signalons la création d'un label d'IA inclusive pour le recrutement créée par **Arborus** et **Bureau Veritas**, ce dernier assurant les audits, évidemment rémunérés[1024].

Gestion des carrières

D'autres applications à base d'IA visent à améliorer la correspondance entre les compétences, le potentiel des collaborateurs et leur parcours dans l'entreprise.

WiserSkills (2016, France) a conçu une solution qui sert à cartographier les compétences à développer dans l'entreprise et à exploiter cette information pour préparer les collaborateurs à s'y adapter. C'est un outil en ligne où le collaborateur décrit lui-même ses compétences et ses sources de motivations. Il ne s'agit pas seulement de renseigner ses seules aptitudes métier. L'outil était en test à la Société Générale en 2017.

Braincities (2013, France) propose une « IA bienveillante » avec des applications dans les RH, la finance ainsi que pour les Smart Cities. Et notamment un outil de machine learning pour analyser les parcours de carrière et le matching entre collaborateurs et équipes dans les métiers techniques. Un autre outil de la startup permet d'analyser les échanges textuels dans l'entreprise pour les associer à ses éléments culturels. Histoire de détecter les comportements déviants ?

Leena AI (2015, Inde) est un chatbot pour RH qui répond aux questions des salariés de l'entreprise. Il s'intègre dans Slack ou Facebook Workplace. Cela couvre surtout les processus standardisés de l'entreprise comme sur sur les congés et les notes de frais.

eightfold.ai (2016, USA, $23,75M) est un outil de gestion de pool de talents basé sur l'exploitation de données internes et externes à l'entreprise. Au bout du compte, tout cela est censé accélérer l'ensemble du processus de recrutement, et qui plus est, de réduire les erreurs de recrutement. Le produit a cependant l'air d'être, comme pour nombre de startups US, adapté au marché américain, ne serait-ce que tu fait des liens avec les logiciels de HR qui sont eux-mêmes nord-américains et pas forcément déployés en France.

HelloElton (2018, France) propose un coach à base d'IA qui forme les collaborateurs aux « soft skills ». Reste à savoir ce que fait exactement cette IA.

[1023] Voir Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices par Manish Raghavan et al, juin 2019 (24 pages) fait un inventaire des méthodes algorithmiques de réduction des biais des IA exploitées dans le recrutement.

[1024] Voir Lancement du 1er label international pour une IA Inclusive par Arborus et Bureau Veritas, septembre 2020. Arborus est un fonds de dotation lancé par l'association Arborus et de grandes entreprises internationales en 2010. Il fait la promotion de l'égalité femmes- les hommes dans le monde, via le le label européen et international GEEIS (Gender Equality European & International Standard).

Seedlink (2013, Pays-Bas) propose une IA qui sert à prédire le comportement des salariés. Elle s'appuie sur du traitement du langage qui analyse le résultat de questions ouvertes d'une application mobile. La solution est utilisée chez L'Oréal (vidéo).

gPulse (2016, France) prend le pouls des salariés avec des enquêtes d'opinion avec questions fermées et ouvertes. Une fois interprétées, elles doivet permettre de « *mesurer la météo sociale, prédire une démission, calculer le salaire d'équilibre, détecter le risque de burn-out et scorer le fit profil/post* ». Cela repose sur du traitement du langage. On demande à voir (*ci-contre*) !

Flashbrand (2015, USA) propose aussi un outil de sondage des collaborateurs mâtiné d'IA qui propose aux managers des questions à poser en fonction des résultats des enquêtes précédentes.

Chez IBM, une IA serait capable de prédire quand les salariés seraient sur le point de démissionner[1025]. En fait, cette IA est utilisée en interne chez IBM. Elle serait efficace dans 95% des cas mais la recette n'est pas précisée. IBM indique avoir réduit de 30% l'effectif de sa RH à l'échelle mondiale. En plus de leur outil de prévision de démissions qui est probablement à base de machine learning multi-critères, la RH propose aux salariés l'assistant virtuel AI MYCA (My Career Advisor) pour identifier les domaines d'amélioration des compétences et Blue Match qui permet d'identifier de nouveaux postes en interne[1026].

Enfin, **Pôle Emploi** utilise sa base de données de 8,5 millions de demandeurs d'emplois (en 2016) pour proposer des parcours de métiers aux demandeurs d'emploi s'appuyant sur leurs compétences existantes et prédire leurs chances de retour à l'emploi[1027]. Ils prédisent aussi les chances qu'un emploi d'un recruteur soit pourvu.

Management

Les solutions d'IA dédiées au management en général sont encore rares. On a vu qu'elles étaient pour l'instant concentrées sur le recrutement et la gestion des carrières. Le management pourrait cependant tirer parti de diverses solutions existantes ou à venir à base d'IA. Cela concerne par exemple celles qui fluidifient le travail collaboratif et qui améliorent la priorisation des tâches. A terme, on verra sans doute apparaître des outils qui permettront aux managers d'améliorer leur communication. C'est encore expérimental à ce stade[1028].

Finance et achats

Les fonctions financières de l'entreprise font peuvent aussi faire largement appel à de l'IA à tous les étages de la fusée. Nous allons examiner le cas de la comptabilité et des achats.

Comptabilité

La comptabilité et l'expertise comptable font appel aux logiciels depuis des décennies. Est-ce l'IA peut impacter ce métier ? Il est à vrai dire transformé de plusieurs manières.

[1025] Voir L'IA d'IBM peut vous dire quand vous allez démissionner ! par Fabrice Auclert, avril 2019.

[1026] Voir IBM's Artificial Intelligence Strategy Is Fantastic, But AI Also Cut 30% Of Its HR Workforce par Dan Pontefract, avril 2019.

[1027] Vu dans L'intelligence artificielle en entreprise du Cigref, octobre 2018 (40 pages).

[1028] Voir What Machine Learning Teaches Us about CEO Leadership Style par Michael Blanding, août 2019 qui fait référence à Machine Learning Approaches to Facial and Text Analysis: Discovering CEO Oral Communication Styles par Prithwiraj Choudhury & Al, 2019 (58 pages). Ce projet de recherché analyse des écrits de 130 CEO du monde entier et leur visage lors de leur communication orale. Sans grande surprise, les CEO ayant une communication positive réussissent mieux que les autres.

La première est la saisie des écritures comptables par les entreprises elles-mêmes, au gré de leur informatisation qui est devenue inévitable même pour les TPE. En gros, le travail de saisie est passé des experts-comptables vers les clients.

La seconde partie du métier relève de la vérification des comptes et des règles comptables, juridiques et fiscales. Elle est réalisée par des experts-comptables puis par des cabinets d'audit. Nombre de vérification de ces règles sont déjà automatisables par les logiciels traditionnels. L'IA peut ajouter son grain de sel en identifiant des bizarerries dans les comptes comme de recettes ou surtout des dépenses ou recettes qui sortent de la normale[1029].

Les solutions de comptabilité servent aussi à gérer sa trésorerie, à la planifier, à créer des business plans en cas d'augmentation ou de modification de la structure de son capital. Elles permettent de préparer les négociations avec les banques en cas de demande de prêts.

Voici un petit échantillon de startups dans ce secteur sachant que je n'ai pas analysé l'offre d'acteurs étalis comme Intuit, Sage, Cegid ou EBP.

Agicap (2016, France) propose une solution de gestion comptable en cloud pour TPE qui s'appuie sur de l'IA pour gérer sa planification de trésorerie, pour savoir quand embaucher, comment gérer son encours client, ses emprunts, etc. Le tout s'appuie sur des techniques de machine learning assez classiques.

iPaidThat (France) est une autre solution de comptabilité qui s'appuie sur de l'IA pour faire du rapprochement de factures et paiements dans la comptabilité. Une IA à base de machine learning, mais il n'est jamais évident de savoir ce qu'elle peut bien faire de concret par rapport aux techniques logicielles habituelles.

Dooer (2015, Suède, $4.6M) ajoute de la reconnaissance d'image, en fait de l'OCR, aux processus comptables, pour scanner et reconnaître le contenu des factures et notes de frais.

Suplari (2016, USA, $13,4M) se positionne sur un domaine voisin de la comptabilité, la gestion des fournisseurs. C'est aussi à base d'IA, non précisée. Dire qu'un tel logiciel s'appuie sur de l'IA est similaire à dire que c'est … du logiciel. Pour être un peu plus spécifique, il faudrait au minimum indiquer quelle technique est utilisée et quelles données servent à entraîner l'IA.

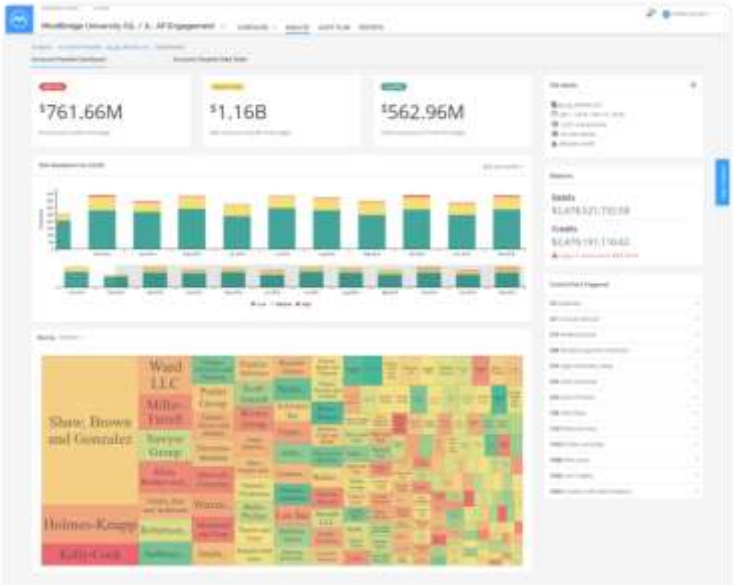
White (2015, France) est une startup qui permet la saisie automatique de pièces comptables pour l'expertise comptable et l'audit. L'outil est capable de comprendre la structure du document et de le traiter convenablement dans son environnement. Il va au-delà des solutions traditionnelles d'OCR (optical characters recognition).

Dhatim (2008, France, 5M€) automatise la gestion des factures et le contrôle des déclarations sociales avec comme premiers clients les opérateurs mobiles (pour les factures), d'autres (pour les déclarations sociales) tout comme le travail de saisie dans les cabinets d'expertise comptable[1030] (vidéo). La solution permet notamment d'automatiser la saisie et l'affectation comptable et d'éviter de générer des incohérences dans les déclarations sociales et les pénalités qui vont avec les contrôles qui sont eux inévitables. La solution s'appuie sur une combinaison de centaines de règles métiers et de machine learning lit les documents non structurées et déclenche des actions automatisées. Elle est notamment utilisée par le réseau d'experts comptables Cogedis.

[1029] Voir L'intelligence artificielle expliquée par l'ordre des experts comptables par Xavier Gaulle, juin 2020 ainsi que Avenir du métier d'expert-comptable : vers une fonction 4.0 propulsée par l'IA et la data visualisation ? par Cyril Degrilart, décembre 2020.

[1030] Voir Dhatim présentera Conciliator Expert, le premier logiciel de saisie comptable complètement automatisé grâce à l'Intelligence Artificielle lors du 73ème congrès de l'Ordre des Experts Comptables, octobre 2018.

Mindbridge AI (2015, Canada, $12,7M) a développé Ai Auditor, une solution d'audit de comptes à base d'IA combinant machine learning, des modèles statistiques et un moteur de règles. Une attention particulière est portée sur l'analyse de l'encours client. Le logiciel permet aussi l'interrogation des comptes en langage naturel (vidéo de démo)[1031].



Georges (France) est un robot expert comptable destiné aux professions libérales et pour 24€ TTC par mois (vidéo). Reste à savoir quelle est la partie humaine et celle qui relève de l'IA dans l'offre.

Des solutions existent depuis longtemps et évoluent pour intégrer des briques d'IA, en général de machine learning, pour identifier des phénomènes anormaux dans les comptes des entreprises. Nous avons de l'automatisation de comptabilité avec **Smacc Hypatos** (2015, Allemagne, $3,9M) qui cible les TPE et PME. Et puis de l'optimisation de planification financière d'entreprises avec **Anaplan** (2006, USA, $300M), **Adaptive Insights** (2003, USA, $22,5M) et **Trufa** (2013, USA, $17M).

Citons enfin l'Américain **H&R Block** qui a mis IBM Watson dans les mains de ses conseillers fiscaux « brick and mortar » pour optimiser la fiscalité de ses clients. C'est un modèle qui sera probablement de plus en plus courant : des AI qui améliorent la productivité des professionnels dans les services mais ceux-ci conservant un contact humain avec les clients. La fiscalité est d'ailleurs à elle-seule un bon champ d'application de l'IA aussi bien du côté des entreprises que du fisc[1032] !

Achats

Les achats peuvent aussi faire appel à des solutions logicielles exploitant des briques d'IA. C'est déjà le cas dans les grandes offres du marché comme chez SAP avec Leonardo et Ariba. L'idée générale des fournisseurs de solution est de réduire la charge de travail des actions répétitives pour libérer du temps utilisable dans l'analyse des données et l'optimisation des achats.

Différentes briques d'IA sont exploitables dans toute la chaîne de valeur des achats et de la finance associée[1033] :

- La **RPA** est mise en œuvre pour automatiser les processus de référencement fournisseurs, achats, gestion des devis, signatures, réception de facturation et réglements fournisseurs, surtout dès lors qu'ils font appel à des applicatifs hétérogènes[1034].
- Les **logiciels d'OCR** servent à scanner et interpréter automatiquement le contenu des factures mais aussi les notes de frais des salariés.

[1031] Ils documentent bien cela dans la présentation How the World's First Auditor Based on Artificial Intelligence Is Driving Change in the Auditing Process, 2018 (16 slides).

[1032] Voir How AI And Robotics Can Change Taxation par Naveen Joshi, janvier 2020.

[1033] Voir quelques autres sources d'information, souvent dilluées, sur les usages de l'IA dans le procurement : The AI revolution in procurement, 2018 (18 pages), Future-proof procurement, KPMG, 2016 (60 pages), Digital Procurement – From Myth to Unleashing the Full Potential, Olivier Wyman (38 pages), Cognitive Procurement: Seizing the AI opportunity, IBM (24 pages) et An SAP Perspective – Procurement 2025, janvier 2018 (24 pages).

[1034] Voir Intelligence Artificielle : quels bénéfices pour les achats ?, avril 2019.

- Des **outils de traitement du langage** peuvent servir à crawler le web pour récupérer des informations pertinentes sur les fournisseurs pour vérifier par exemple certains éléments de conformité. Ils peuvent aussi servir tout simplement à trouver de nouveaux fournisseurs en fonction de critères de recherche multiples. A l'instar des nombreuses solutions à base d'IA du monde juridique, le traitement du langage peut aussi accélérer la gestion des contrats avec les fournisseurs[1035].

- Des **moteurs de règles** « à l'ancienne » sont aussi employés dans les solutions pour gérer les clauses d'achats, le respect de la législation, les remises, etc.
- Les **outils d'analytics** à base de machine learning peuvent servir à analyser les informations historiques et à identifier les anomalies comme des écarts de prix entre contrats, tarifs pratiqués en ligne et devis.
- Les **chatbots textuels** peuvent aussi aider les personnels des services achats à trouver rapidement les bonnes informations sur les fournisseurs. Certains comme EDF ont déjà mis en production un chatbot gérant un premier niveau de réponse aux requêtes des fournisseurs.

Les acteurs de ce marché sont dans l'ensemble plutôt des éditeurs de logiciels historiques qui ont adopté les technologies de l'IA au fil de l'eau. Les entreprises peuvent aussi développer leurs propres solutions sur mesure s'ils sont très spécifiques. Dans tous les cas de figure, la qualité et l'homogénéité des données de gestion est critique pour déployer des solutions d'IA opérationnelles.

[1035] Source du schéma : Artificial Intelligence and its impact on procurement and supply chain par GEP, 2018 9 (55 pages).

ScoutBee (2015, Allemagne, $76M) propose une solution de recherché de fournisseurs qui exploite du machine learning et du traitement du langage pour faire donc du « Smarter Supplier Discovery ».

Jaggaer / SciQuest (1995, USA) est un éditeur de logiciel en cloud établi qui gère toute la filière achats avec des outils pour le sourcing de fournisseurs, la gestion de contrats, le suivi des dépenses, la gestion d'inventaire et de stocks et la gestion de workflow complète. Et bien évidemment, ils ont saupoudré leur offre de diverses briques d'IA. Elle cible surtout des clients dans l'éducation, la santé et le secteur public. La société est présente en France via l'acquisition d'une startup américaine qui était déjà présente en France, **BravoSolution** (1995, USA). Jaggaer a aussi fait l'acquisition de **Pool4Tool** (2000, Autriche) qui a créé une plateforme collaborative entre fournisseurs et clients.

Zycus (1998, USA) est un éditeur de logiciel de systèmes de gestion des achats généraliste positionné au départ sur l'aval des processus achat comme l'identification d'économies.

Ivalua (2000, USA, $134M) est un autre acteur généraliste et international de ce secteur.

Basware (2002, USA) est un opérateur de dématérialisation et de gestion des achats qui détecte les factures falsifiées ou erronées via du machine learning. Ils font aussi de l'OCR de documents de la détection prédictive de fraudes.

FairMarkit (2017, USA, $41,9M) est un autre acteur de l'IA du procurement qui s'intéresse à la longue traîne des dépenses de moins de $1M. Leur plateforme agrège l'offre des fournisseurs et joue un rôle de place de marché et de gestionnaire d'appels d'offres pour trouver les mieux disants. Cette gestion est alimentée par des briques de machine learning qui analysent l'historique des achats pour faire de la recommandation de fournisseurs déjà référencés ou non référencés. Les fonctions de recherche de produits font aussi appel à du machine learning pour le traitement du langage.

Applications métiers de l'IA

Nous allons faire un tour d'horizon des usages de l'intelligence artificielle dans un bon nombre de marchés verticaux dans cette grande partie de plus 190 pages

Tous les secteurs d'activité ont adopté progessivement des solutions à base d'IA qu'elles soient génériques comme celles que l'on trouve dans nos smartphones et micro-ordinateurs ou qu'elles soient adaptées aux métiers. Leur rythme d'adoption de l'IA dépend de paramètres comme la densité d'innovations dans leur domaine, les quantités de données gérées le nombre de startups du domaine, la santé économique du secteur et sa fragmentation soit internationale soit nationale qui peut impacter la vitesse de propagation des innovations. Les secteurs d'activité exploitant le plus l'IA sont ceux qui génèrent le plus de données, comme ceux de la finance, des transports, de la distribution ou de la santé. Le moins outillé est celui de l'éducation, du fait de sa fragmentation très élevée et de la difficulté à tenir les promesses un peu trop optimistes des IA qui accompagnent le parcours des élèves de manière personnalisée.

Les métiers de la relation client exploitent surtout les outils de traitement du langage. Ceux qui exploitent des données chiffrées comme la finance font beaucoup appel à du machine learning. Les activités ayant un lien avec le monde physique comme l'industrie manufacturière ou l'agriculture utilisent principalement l'analyse de données issues d'objets connectés, la vision artificielle et la robotique.

transports

santé

manufacturing

finance

assurance

agriculture

utilities

construction

distribution

tourisme

juridique

médias

éducation

services publics

défense et sécurité

L'inventaire de cette partie couvre divers projets clients. Il est surtout alimenté par les offres de startups provenant du monde entier. Elles ne sont pas toujours bien documentées, notamment d'un point de vue quantitatif, ce d'autant plus, que nombre d'annonces portent sur des « proof of concepts » et pas forcément sur des solutions déployées à grande échelle. Elles fournissent tout de même une bonne indication des usages et des tendances.

Transports

Le marché des transports est vaste avec le transport routier, aérien, ferroviaire, fluvial et maritime.

Tous ces secteurs sont transformés de près ou de loin par l'IA. Par exemple, sans que cela transparaisse, les systèmes d'optimisation du transport maritime par containers sont de plus en plus optimisés par des techniques de machine learning, voire de deep learning, remplaçant des techniques tradi-

tionnelles. Le *yield management* des compagnies aériennes bénéficie aussi de ces avancées en intégrant a minima du machine learning à défaut de deep learning.

C'est surtout le transport automobile qui progresse le plus grâce à l'IA, via les véhicules à conduite assisté ou autonome que nous allons examiner de près. Ainsi que l'interconnexion entre les transports et leur optimisation pour permettre aux villes de résoudre les problèmes de congestion et de pollution. Ces évolutions seront d'ailleurs accélérées avec l'arrivée des véhicules autonomes qui n'auront plus vraiment de lignes de dessertes comme les bus pour ceux qui seront partagés.

Route

L'intelligence artificielle intervient à se nombreux endroits pour ce qui est de la route. On pense bien entendu à la conduite assistée et autonome, que nous allons examiner en détail, mais ce n'est pas le seul. Les parkings, les péages, d'autres parties des véhicules que la partie conduite pure peuvent aussi faire appel à de l'IA.

Conduite assistée et autonome

Les véhicules autonomes constituent sans conteste un défi technologique difficile à relever tout comme un facteur de changement énorme pour la vie de tous les jours et pour un grand nombre de secteurs d'activités, dans la ville intelligente, la construction, les télécommunications, les contenus et les assurances pour ne retenir que les plus évidents.

Techniquement parlant, les véhicules autonomes sont des robots capables d'atteindre un objectif en tenant compte de leur environnement et d'imprévus. À la différence des robots humanoïdes, ils sont cependant bien plus matures, même s'ils ont fort à faire pour interagir avec les Humains. La raison est simple : même si c'est une tâche complexe, faire rouler un véhicule sur une route, malgré toutes les contraintes que cela représente, présente moins de difficultés que de se mouvoir dans l'espace en 3D et d'interagir avec l'environnement physique.

Dans une voiture, la surface de contact est relativement simple et limitée : un plan et des roues ! Plus est faible le nombre de degrés de liberté, plus l'automatisation est facile à gérer. C'est pour cela que les métros automatiques comme les lignes 1 et 14 à Paris sont déjà monnaie courante ou que les avions volent le plus souvent en pilote automatique, sauf lors des phases de décollage et d'atterrissage qui pourraient d'ailleurs être très souvent également gérées avec le pilote automatique.

Cela explique aussi pourquoi des minibus autonomes comme ceux des français **Navya** et **EasyMile** peuvent circuler dans certains lieux publics où les autres véhicules ne circulent pas[1036].

L'hétérogénéité des véhicules circulant génère une complexité que les véhicules autonomes doivent gérer et en particulier lorsqu'ils doivent interagir et réagir aux passants, cyclistes et autres engins pilotés par des humains. Plus c'est hétérogène, plus c'est complexe.

Navya
Arma
autonome

EasyMile
EZ10
autonome

Local Motors Olli
imprimé en 3D, dialogue avec passagers géré avec IBM Watson IoT

[1036] Les véhicules de **Navya** sont opérationnels dans des lieux protégés en Australie, Nouvelle Zélande, à Taiwan, Singapour, au Canada, au Luxembourg, au Danemark, en Autrice, en Suisse et à l'aéroport d'Heathrow à Londres. Aux USA, ils sont déployés à Las Vegas et doivent l'être à Disney World en Floride. En France, on en trouve à Paris, Lyon et Nantes. Les **EasyMile** EZ10 (2014, France, $22,1M) sont testés en Californie, Australie, à Taiwan, en Estonie, aux Pays-Bas, en Estonie et en Allemagne. Il faut compter aussi avec les Olli de **Local Motors** (2007, USA, $250K) ainsi qu'avec les K4 L4 Appolong du constructeur **King Long** qui sont équipées du système d'exploitation de véhicule autonome Apollo OS de **Baidu** et roulent jusqu'à 70 km/h.

D'où l'idée de mener des tests avec 100% de véhicules autonomes[1037] !

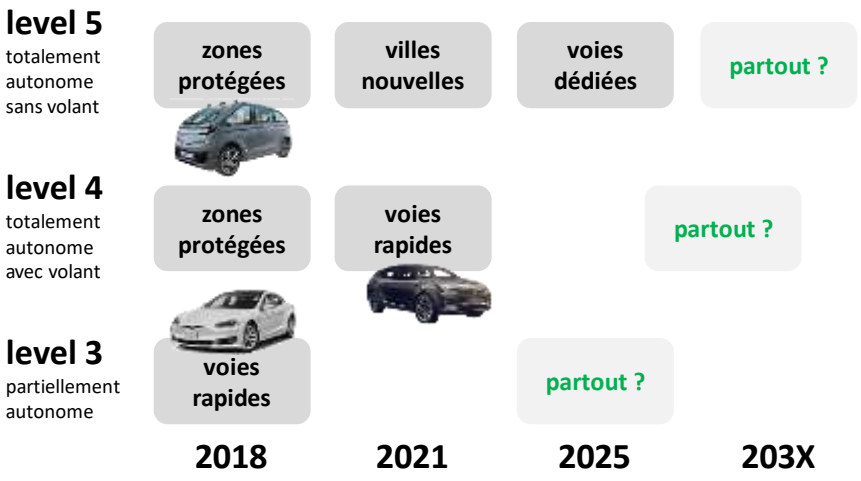
Les véhicules autonomes sont une « réalité progressive » avec des niveaux d'autonomie étalés entre le niveau 3 (« sans les mains», le niveau 4 (« sans les yeux») et le niveau 5 (« sans la tête et l'esprit tranquille »). Elle existe. Elle est démontrée.

Si elle n'est pas encore courante, son contexte d'utilisation crédible s'agrandit d'année en année.

Les constructeurs et équipementiers inventent aussi des niveaux de conduite assistée semi-autonome intermédiaire comme le L2+ qui supporte par exemple l'ajustement automatique de la vitesse, la sortie d'autoroute, etc. Ce niveau d'autonomie est notamment promu par l'équipementier allemand **ZF** ainsi que par **Mobileye**.

On passera très graduellement de l'autoroute en conduite semi-autonome à la conduite en route traditionnelle, puis en dernier ressort, en ville. Elle méritera alors pleinement son appellation d'automobile !

La phase la plus délicate sera d'intégrer la conduite autonome dans des villes embouteillées et surtout hors des USA (Naples, Calcutta, Shanghaï, la place de l'Etoile à Paris) et à les faire cohabiter avec des véhicules à conduite traditionnelle, sans compter les deux roues et les piétons[1038]. Le problème sera techniquement plus simple à gérer lorsque 100% des véhicules seront à conduite automatique dans les villes.



[1037] Cela aurait du sens de lancer un appel d'offre pour des villes de taille intermédiaire pour y expérimenter une conduite 100% autonome, avec des partenaires industriels. Il faudrait pour cela disposer aussi de véhicules utilitaires autonomes et, pourquoi pas, revoir la structure de la voirie. Il est probable que cela sera un jour expérimenté en Chine ou aux USA. En attendant, on peut compter sur l'expérience de **Transpolis** qui recréé les conditions d'usage de véhicules autonomes dans une zone dédiée à l'Est de Lyon et au Nord de l'Aéroport Saint Exupéry. Transpolis est une *joint venture* associant divers industriels dont Renault Trucks, Colas, Vibratec, Groupama et la Caisse des Dépôts. La Chine va carrément construire de toutes pièces une ville de plus de 2 millions d'habitants sur 100 km^2, équipée des dernières technologies, dont des véhicules autonomes, Xiong'an New Area, à 90 km à l'Est de Beijing. Le gouvernement français annonçait sinon en avril 2019 lancer un programme d'expérimentation de conduite autonome avec 16 expériences montées par deux consortiums. L'État finance 42M€ sur les 120M€ des projets, complété par un apport des régions..

[1038] Voir ce projet de recherche de prévisions du mouvement des piétons avec Bio-LSTM: A Biomechanically Inspired Recurrent Neural Network for 3-D Pedestrian Pose and Gait Prediction, 2019 (vidéo).

Pour régler les problèmes d'hétérogénéité, on interdira à long terme la conduite manuelle. C'est le stade 5 de la conduite autonome. Enfin, c'est une hypothèse[1039] !

Les progrès de ces dernières années résultent d'efforts qui ont démarré en 2002 lors de défis lancés aux chercheurs américains par la DARPA[1040]. Celui-ci consistait à faire rouler en mode autonome des véhicules sur un parcours de 240 km dans le désert du Mojave entre la Californie et le Nevada et en touchant en premier la ligne d'arrivée. Avec à la clé, un prix de $1M. Le premier défi fut organisé le 13 mars 2004. Aucun véhicule ne put faire la course en entier. Le véhicule de la Red Team de l'Université Carnegie Mellon roula juste moins de 12 km.

Il fallut attendre le second défi du 9 octobre 2005 pour voir celui-ci relevé avec succès, par l'équipe de **Stanford**[1041], suivie de celle de **Carnegie Mellon** associée à l'**Université de Pittsburgh** en Pennsylvanie[1042]. 40 équipes avaient été sélectionnées en tout pour cette course qui faisait 212 km, également en Californie.

Un troisième défi fut organisé en 2007 avec $2M de prix. Il s'agissait de rouler en milieu urbain.

Le gagnant fur l'équipe Tartan Racing associant l'Université de Carnegie Mellon University et General Motors avec Boss, une Chevrolet Tahoe modifiée[1043] (slide *ci-dessus*). Le second de la course fut une équipe de Stanford utilisant une Volkswagen Passat modifiée.

Dans les contraintes à respecter, il fallait suivre le code de la route californien, être entièrement autonome, la route à suivre n'était fournie par la DARPA que 24 heures avant la course dont il fallait respecter les étapes, les véhicules ne pouvaient pas s'arrêter plus de 10 secondes, devoir pouvoir rouler sous la pluie et dans le brouillard, sans GPS, éviter les collisions de toute nature, pouvoir circuler dans un parking et faire un demi-tour.

Dans les dix années qui suivirent, les progrès furent plus laborieux et incrémentaux. **Google** lançait son projet de véhicule entièrement autonome vers 2010[1044], un véhicule sans volant de niveau 5. Depuis, la filiale Waymo d'Alphabet, la maison mère de Google, fait l'acquisition de véhicules autonomes pour déployer une flotte de taxis autonomes aux USA avec 20 000 SUV I-Pace chez **Jaguar** en mars 2018 et 62 000 chez **Fiat Chrysler** en mai 2018.

[1039] Philippe Méda décrit ce scénario dans The 5 perimeters of autonomous vehicles, novembre 2019. La question n'est plus de savoir quand les véhicules autonomes circuleront mais à quelle vitesse leur champ d'action s'élargira des zones protégées à des zones qui le sont moins.

[1040] L'histoire est racontée dans Inside the races that jump-started the self-driving car de Alex Davies, 2017 et The Races That Jump-Started the Self-Driving Car de Wired, 2017 (14 minutes). Voir aussi la video The great robot race (52 minutes).

[1041] Dans l'équipe des gagnants, on trouve notamment Joshua Anhalt, maintenant chez Uber, Hong Bae chez Faraday Future (un constructeur Chinois filiale de LeEco vu aux CES 2016 et 2017), Chistropher Baker, dans projet iSee du MIT et d'autres encore qui sont pour la plupart dans l'écosystème des véhicules autonomes.

[1042] Il fallut attendre 2017 pour comprendre pourquoi l'équipe de Carnegie Mellon avait échoué et obtenu que la seconde place en 2005. Il s'agissait d'un problème de filtre dans le moteur, qui n'avait rien à voir avec la fonction de conduite autonome. Voir Carnegie Mellon Solves 12-Year-Old DARPA Grand Challenge Mystery - Highlander narrowly took second in the 2005 DARPA Grand Challenge. Just last week, CMU figured out why de Evan Ackerman, octobre 2017. De son côté, l'équipe de Cornell avaité été pénalisée par un GPS défaillant.

[1043] Leur performance est documentée dans Autonomous Driving in Urban Environments: Boss and the Urban Challenge, 2008 (42 pages).

[1044] Voir le TED de Sebastian Thrun Google's driverless car, 2011.

En 2015, **Tesla** lançait la fonction Autopilot de niveau 3 pour rester dans sa file d'autoroute. Elle s'appuyait sur une caméra et un système embarqué de l'israélien Mobileye, sans LiDAR.

En 2016, Tesla**,** démontrait que ses voitures autonomes pouvaient faire un trajet complet de manière automatique, au-delà des fonctionnalités de l'Autopilot qui est surtout censé servir à rester dans sa voie sur autoroute (conduite automatique de niveau 3). Voir la vidéo 1 et la vidéo 2 avec une Tesla X. Certes, les rues empruntées ont un trafic très faible, elles sont très larges et aucun piéton n'est visible, comme souvent aux USA. Des situations que l'on rencontre plus rarement dans les villes européennes. Les démonstrations des Google Car sont du même acabit même si elles circulent plus lentement que les Tesla.

En 2019, Elon Musk continuait la surenchère en annonçant que la conduite entièrement autonome serait disponible en 2020 par mise à jour de sa base installée. Nous sommes en 2021 et ce n'est toujours pas le cas. Même si Tesla fait des progrès dans le logiciel et autour de sa propre architecture matérielle embarquée exploitant son processeur maison[1045]. Ce qui au passage sous-entend que la promesse concernant la base installée est surjouée. Elon Musk décrivait cependant le modèle économique à venir des véhicules autonomes, pouvant devenir des robot-taxis autonomes lorsque l'on n'en a pas usage.

L'histoire se répète chaque année et il faut toujours prendre cela avec un gros grain de sel, tout comme les dates de livraison promises pour le Semi (*ci-contre*) annoncé en 2017 [1046] ou le CyberTruck lancé en 2019, censés également devenir autonomes[1047].

Audi commercialise depuis 2018 ses A8 motorisées avec un V6, disposant de conduite semi-autonome de niveau 3 avec encore plus d'autonomie que l'Autopilot de Tesla, notamment à basse vitesse.

Mais elle n'est pas encore autorisée avec ce niveau d'autonomie aux USA.

En 2015 et 2016, plusieurs expériences de conduite autonome de camions ont été réalisées en Europe, avec notamment **Volvo**. Des milliers de kilomètres ont été parcourus par une série de camions sur des voies rapides traversant plusieurs pays.

Il faut aussi creuser derrière les effets d'annonce. Ainsi, **Uber** annonçait lancer son premier service pilote de voitures autonomes à Pittsburgh en septembre 2016 avec des **Ford** Fusion. Mais les véhicules étaient tout de même pilotés, ou tout du moins contrôlés, par des conducteurs dans un premier temps !

Une expérience menée à San Francisco avec 16 véhicules de tests **Volvo** XC90 PHEV a ensuite tourné court fin 2016 après une interdiction par la municipalité de la ville. Uber a alors déplacé ses véhicules en Arizona, plus accueillant.

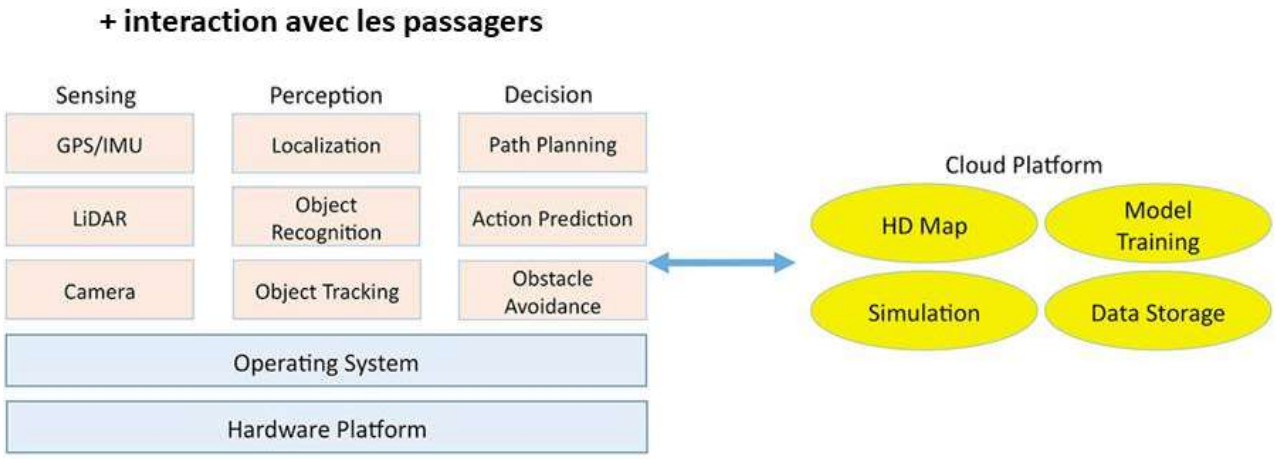
[1045] Voir Tesla Dojo Supercomputer Explained — How To Make Full Self-Driving AI par Maarten Vinkhuyzen, novembre 2020, Elon Musk is annoying the hell out of people who work with self-driving cars par Sasha Lekach, 2019, Tesla seul en tête dans la course à l'autonomie par Frédéric Charles, avril 2019, et Elon Musk says Tesla robotaxis are coming — and soon par Sasha Lekach, 2019.

[1046] Tesla faisait l'annonce de son Semi en novembre 2017 mais ils ne seront pas commercialisés avant longtemps au rythme où vont les choses chez ce constructeur. Les précommandes étaient toutefois lancées avec une petite avance de $15K. Il comprend une cabine pour le livreur !

[1047] Voir Watch Tesla's 'Full Self-Driving' mode fail on San Francisco's twistiest street par Sasha Lekach, octobre 2020.

En 2019, **Volvo** et **Uber** revoyaient leur approche Au lieu de modifier des véhicules existants, la production des nouveaux XC90 sera intégrée dès le départ en usine. L'autonomie est censée passer au niveau 5[1048]. Les effets d'annonce ont aussi porté sur les camions autonomes. **Volvo** faisait déjà des tests routiers en 2016 en Europe.

Uber démarrait de son côté des tests de camions autonomes en mars 2018, issus de l'acquisition de la startup **Otto** (2016, USA), l'acquisition ayant eu lieu la même année que sa création, et relevant d'une « acquihire »[1049]. Ces tests portaient sur l'autonomie de la conduite sur longue distance, les conducteurs reprenant le contrôle des véhicules hors des autoroutes en en ville. Tout ça pour abandonner la partie en juillet 2018 et se rabattre exclusivement sur les voitures autonomes, abandonnées à leur tour fin 2020. Les créateurs d'Otto ont ensuite créé **Kodiak**, un fabricant de camions autonomes.



Un grand nombre de techniques sont mises en œuvre dans les véhicules à conduite assistée ou autonome : de nombreux capteurs (ultra-sons, radar, vidéo, LiDAR), des processeurs dédiés aux systèmes de vision artificielle temps réel (Mobileye, Nvidia, …), des télécommunications (la 5G jouera un rôle clé dans la communication entre véhicules, et entre véhicules et infrastructures connectées), des services en cloud (cartographie 2D et 3D des environnements, cartographie pour déterminer la route) et des systèmes experts de prise de décision.

Les progrès récents sont dus aux avancées parallèles dans tous ces domaines. L'écosystème qui se met en place fait intervenir de nombreux acteurs spécialisés et créant des produits qui deviennent des plateformes comme les processeurs de Nvidia.

Équipementiers

Pour comprendre son environnement, un véhicule autonome doit disposer d'une vision stéréoscopique ou 3D. C'est aujourd'hui le rôle des LiDAR avec leur laser tournant mais ils sont pour l'instant trop chers, coûtant plusieurs milliers d'Euros l'unité. Leur marché est dominé par le californien **Velodyne** et quelques copycats chinois.

[1048] Voir Volvo a travaillé sur la sécurité de la nouvelle voiture autonome d'Uber par Nicolas Furno, juin 2019.

[1049] Otto avait été cofondé par Anthony Levandowski, un ancien de l'équipe Waymo de Google chez qui il était entré en 2007 après avoir participé au projet de deux roues autonome Ghostrider de l'Université de Berkeley qui avait concourru au Grand Challenge de la DARPA en 2004 et 2005. La cofondatrice d'Otto était la Française Claire Delaunay, qui était aussi passée par Google et est maintenant VP Engineering chez Nvidia.

La bataille en cours consiste à créer des LiDAR dits « solid state » n'ayant pas besoin de pièces mécaniques mobiles rotatives contrairement aux premiers LiDARs apparus sur le marché. Nombre de startups comme **Quanergy** (2013, USA, $135M) ou **LeddarTech** (2007, Canada, $123M) proposent ce genre de solution, mais avec des angles de vue limités qui obligent à cumuler plusieurs capteurs pour disposer d'un angle de vue équivalent aux LiDAR rotatifs.

L'autre solution consiste à utiliser des capteurs 2D traditionnels et du deep learning pour interpréter les scènes. C'est ce que fait **Mobileye** (1999, Israël, $515M) qui s'est fait gober par Intel pour $15,3B en 2017[1050]. Enfin, **Outsight** (France) développe une sorte de LiDAR multi-spectral fonctionnant à la fois dans le visible et l'infrarouge, couplé à un logiciel d'analyse et de classification. Il permet de cartographier l'environnement avec précision, y compris la vitesse des objets détectés.

Les systèmes embarqués intègrent le plus souvent un GPU Nvidia adapté au deep learning d'interprétation des images générés par ces capteurs divers. La résolution des images traitées par ces systèmes est encore médiocre, ce qui limite leur précision, mais suffit aux usages actuels. Elle s'améliorera sans doute avec les progrès à venir de ces GPU et autres processeurs neuromorphiques.

Tesla est probablement le constructeur qui a le plus de véhicules semi-autonomes en circulation avec ses Model S et le mode Autopilot qui est régulièrement mis à jour. La fonction Autopilot qui a des équivalents chez d'autres constructeurs n'est pas la seule qui automatise certaines tâches de la conduite.

[1050] Voir Engines Powering L2+ to L4 de Mobileye, janvier 2020 (64 slides) qui décrit en détail les enjeux technologiques de la conduite avec un niveau d'autonomie L2+.

On peut aussi compter sur :

- Le **parking automatique**, comme avec Park4U de Valeo qui est installé sur de nombreux véhicules haut de gamme de marques allemandes et françaises. Fini les créneaux difficiles à réaliser !
- Le **maintien dans sa file** sur voie rapide (Lane Keeping Agents), une des fonctions clés de l'autopilot des Tesla S.
- Les **manœuvres** avec des agents capables de doubler un véhicule et d'autres qui permettent de sortir de la voie rapide. Il existe aussi des agents qui évitent les collisions.

Les équipementiers et les constructeurs se mettent tous en branle pour se préparer à cette évolution radicale du marché. Ils sentent le vent venir et veulent aussi éviter de se faire dépasser par les acteurs issus du numérique tels que Google.

C'est le cas de **Ford** qui a lancé en 2017 sa filiale commune avec Carnegie Mellon, Argo AI, dotée d'un budget de $1B. Ce même Carnegie Mellon gagnant ou second des DARPA Grand Challenges ! En 2019, Argo AI en était à sa troisième génération de véhicules autonomes construits sur une base Ford Fusion Hybrid. Elle utilise une nouvelle génération de capteurs, aussi bien au niveau des radars que des caméras à plus haute révolution et dynamique.

Des tests sont réalisés à Pittsburgh, Austin, Detroit, Miami, Palo Alto et Washington, D.C[1051]. De son côté, Toyota lançait en 2019 un fonds d'investissement de $100M dans les startups de l'IA[1052].

Waymo est la branche véhicule autonomes d'Alphabet. Elle travaille notamment avec **DeepMind** sur un système dénommé Population Based Training (PBT) qui est meilleur pour interpréter l'environnement. Il utilise un algorithme évolutionnaire qui rappelle celui de PathNet[1053].

Renault Nissan s'est aussi lancé, l'annonçant au CES 2017. Un test de déploiement de Zoe électrique autonome est lancé à Rouen avec l'opérateur **Transdev**[1054].

PSA n'est pas en reste, ayant déjà testé en France une Peugeot autonome en 2015 et un C4 Picasso en juin 2017. Dans ce cadre, le constructeur français s'est associé à **nuTonomy** (2013, USA, $19,6M) qui développe les logiciels de pilotage également utilisés par Ford ainsi qu'avec **Inria** en 2018 avec l'ouverture d'un Openlab avec l'Institut PRAIRIE (PaRis Artificial Intelligence Research InstitutE). En 2019, PSA annonçait cependant laisser tomber la R&D sur les niveaux 4/5.

En Chine, le Uber local **DiDi** lançait en août 2019 une spin-off dédiée à la conduite autonome. Elle financée à hauteur de $500M par Softbank et employait déjà 400 personnes fin 2020. Elle ambitionne de lancer un service de taxis autonomes. Mais la filiale n'a pas encore de nom (en janvier 2021).

En avril 2019, le gouvernement français annonçait le lancement de 16 expérimentations de véhicules autonomes sur des routes variées (rurales, zones urbaines, etc) et pour tous types de transport (passagers, marchandises)[1055].

[1051] Voir Argo AI is taking its new generation of self-driving vehicles into downtown Detroit par Kirsten Korosec, juin 2019.

[1052] Voir Toyota AI Ventures launches $100M fund to invest in robotics and autonomous tech par Kirsten Korosec, mai 2019.

[1053] Voir Waymo and DeepMind mimic evolution to develop a new, better way to train self-driving AI par Darrell Etherington, juillet 2019.

[1054] Voir Rouen Normandy Autonomous Lab avec Transdev et Renault Zoé, juin 2018.

[1055] Voir Le gouvernement lance 16 expérimentations de voitures autonomes sur les routes de France par Kozi Pastakia, 2019. C'était le résultat d'un appel à projets lancé en 2018 et doté de 200 Md€.

Voici quelques autres startups du secteur automobile, qu'il est très difficile de départager. Ils utilisent généralement des bases technologiques voisines :

- **Optimus Ride** (2015, USA, $23,3M) une spin-off du Massachusetts Institute of Technology qui développe la partie logiciel de véhicules autonomes de niveau 4, y compris de chariots élévateurs (vidéo).
- **Zoox** (2014, USA, $1B), filiale d'Amazon, présentait en décembre 2020 son premier taxi entièrement autonome (level 5). Sa batterie lui donnerait 16h d'autonomie (ce qui est étonnant). Il roule avec une vitesse de pointe de 120 km/h. Il peut transporter quatre passagers[1056]. Il ressemble de près à un Navya (vidéo) !
- **AutoX Technologies** (2016, Chine/USA, $160M), financé à hauteur de $100M par le Chinois Dongfeng Motors en septembre 2019, a déjà commencé à déployer une flotte de taxis autonomes en Chine, une centaine dans plusieurs villes en 2020, dont Shenzhen. La startup a développé sa propre architecture matérielle embarquée, la XCU qui est intégrée dans des Chrysler Pacifica modifiées. Le tout est couplé à du sensor fusion, de la cartographie 3D des villes et un service en cloud (vidéo).
- **Pony.ai** (2016, Chine/USA, $993M) ambitionne aussi de développer un système complet de conduite autonome, un de plus. La startup a été en partie financée par Toyota. Elle fournirait déjà Hyundai en systèmes de conduite autonome qui sont expérimentés en Californie.
- **Netradyne** (2015, USA, $16M) est un spécialiste de deep learning appliqué à la vision des véhicules avec leur plateforme matérielle et logicielle Driveri qui s'installe sur des véhicules existants pour de la conduite assistée (vidéo).
- **Drive.ai** (2015, USA, $77M), créé par des anciens de Stanford, propose aussi une plateforme de conduite autonome à base de deep learning. Ils lançaient une expérimentation de véhicules autonomes au Texas en mai 2018[1057] (vidéo) qui ont la particularité d'intégrer un panneau lumineux de chaque côté pour communiquer textuellement avec les piétons pour leur indiquer s'ils peuvent ou pas passer devant le véhicule.

- **Metamoto** (2016, USA, $2M) propose un système de simulation numérique du comportement de véhicules autonomes dans un environnement de circulation réaliste. **Cognata** (2016, Israël, $23,6M) fait de même avec son outil de simulation pour modèles virtuels de véhicules autonomes.
- **Humanising Autonomy** (2017, UK, $6M) est spécialisé dans la prédiction du comportement des piétons. La solution agrège des techniques issues de la psychologie comportementale, d'IA probabilistes et de techniques de deep learning. Leur solution peut s'intégrer dans des systèmes d'assistance à la conduite de conducteurs humains. Elle peut aussi s'intégrer dans des systèmes de ville intelligente connectées exploitant de la vidéo surveillance et reliés aux véhicules en mode V2X.

[1056] Voir Zoox unveils a self-driving car that could become Amazon's first robotaxi par Sean O'Kane, décembre 2020.

[1057] Voir Drive.ai is launching an autonomous ride-hailing network in Texas de Megan Rose Dickey, mai 2018.

- **Latent Logic** (2017, USA, $2,9M) était une spin-off de l'Université d'Oxford qui développait un simulateur de conduite autonome simulant le comportement de piétons, deux roues et autres véhicules. Elle a été acquise par Waymo, la filiale d'Alphabet, en décembre 2019.
- **AImotive** (2015, Hongrie, $67,7M) développe une famille de logiciels liée à la conduite assistée et autonome avec aiSim, un simulateur de conduite autonome servant à la mise au point des systèmes, aiDrive qui gère cette conduite et la perception sensorielle associée et aiWare, une architecture matérielle supportant aiDrive avec un chipset dédié multicœurs de 100 TOPS, tournant à 1,5 GHz et gravé en 16nm.
- **Comma.ai** (2015, USA, $8,1M) ambitionne de proposer une sorte de SDK permettant de rendre autonomes des véhicules existants avec leur OpenPilot (vidéo). Le logiciel associé est open source. Il exploite un smartphone qui surveille le conducteur pour vérifier qu'il est bien au contrôle de son véhicule.
- **Cortica.ai** (2007, Israël, $69M) se veut le champion de l'IA qui apprend toute seule, par apprentissage non supervisé et par renforcement, avec leur "autonomous AI" et leurs 200 brevets associés. Ils interviennent notamment dans la vision artificielle appliquée à la conduite autonome et notamment pour anticiper les mouvements des objets et personnes sur la route (vidéo). Leur IA est aussi adaptée au scanning de bagages et à l'exploitation de vidéosurveillance (vidéo). Mais il est bien difficile de faire la part des choses entre leur discours et les démonstrations qui semblent exploiter des réseaux de neurones convolutifs qui ont l'air d'être assez classiques pour la détection et la classification d'objets.
- **Predina** (2016, UK) propose des modèles prédictifs pour les véhicules autonomes ou semi-autonomes afin de limiter les risques d'accidents. Le machine learning exploite un historique d'accidents et de collisions pour déterminer leurs causes et prévenir les accidents en incitant les conducteurs à être prudents aux bons endroits.
- Le laboratoire CSAIL et **MIT** expérimente un système permettant aux véhicules autonomes de bien voir la route par temps neigeux ou dans le brouillard en couplant des capteurs à pénétration de sol habituellement utilisés dans le BTP et de données cartographiques[1058].
- **Mindtronic AI** (2018, Taiwan/USA) développe un système d'aide à la conduite qui présente la particularité d'appliquer le principe inverse de la vigilance du conducteur sur la conduite assistée. Il peut reprendre le contrôle du véhicule s'il estime que le conducteur s'y prend mal pour une raison ou pour une autre, par exemple, s'il détecte que le conducteur d'assoupit, est fatigué ou commet des erreurs de conduite manifestes[1059]. Le système s'appuie surtout une caméra embarquée qui surveille le conducteur. Nombre de startups proposent des systèmes de ce genre mais qui se contentent d'alerter le conducteur s'il s'endort.
- **Ween** (2014, France, $1,8M) a développé une IA qui détecte les mouvements des utilisateurs dans leurs usages quotidiens et dans leurs trajets avec une solution à basse consommation. Après un apprentissage par renforcement, cette solution prévoit en temps réel l'arrivée des utilisateurs à leur prochaine destination (domicile, voiture, bureau, transport public...), ce qui permet de proposer un accueil personnalisé au domicile (mise en température du logement, éclairage bureaux, etc).
- **Agility Robotics** (2015, USA, $8M) développe pour Ford le Digit, un robot humanoïde censé faire des livraisons à partir de véhicules autonomes (vidéo).

[1058] Voir To self-drive in the snow, look under the road par Adam Conner-Simons, février 2020.

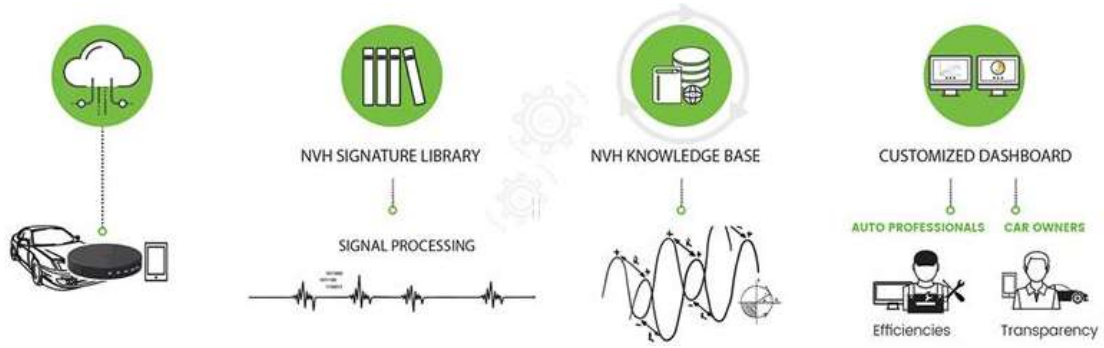
[1059] Voir This AI-Powered Cockpit Knows When To Cut Off The Driver par Ralph Jennings, janvier 2020.

- **Deep Glint** (2013, Chine, $23M) propose des technologies de reconnaissance faciale ainsi que d'analyse de comportement. Ils sont partenaires du Coréen Hyudai qui veut intégrer cette offre pour fournir un service de recommandation automatique de musique et d'éclairage d'ambiance en fonction de l'état du conducteur et des passagers. L'autre application évidente est la détection de l'assoupissement du conducteur, mais ce marché est déjà occupé.

- **Helm.ai** (2016, USA, $13M) commercialise des logiciels de gestion de véhicules autonomes allant du niveau L2 au L4 qui couvrent la perception visuelle (segmentation de scène, sensor fusion avec données de LiDARs, suivi du marquage au sol, détection de piétons), l'analyse d'intentions (véhicules, piétons, deux roues), la planification de la conduite et le contrôle du véhicule.

En février 2018, **Huawei** et **Baidu** collaboraient pour permettre à un smartphone Mate Pro 10 de piloter un véhicule autonome. L'idée était d'utiliser la caméra du smartphone et le NPU du chipset Kirin 970 (processeur pour le deep learning) pour reconnaître l'environnement du véhicule et éviter les obstacles. C'était testé sur une **Porsche** Panamera modifiée. Le tout s'appuyait côté logiciels sur la plateforme HiAI de Huawei (SDK mobile pour le Kirin 970), PaddlePaddle de Baidu (SDK de deep learning) et Baidu Brain (vidéo).

Enfin, citons **Apple** qui est enfin sorti du bois en 2018 pour expérimenter relativement discrètement un logiciel de véhicule autonome dans une cinquantaine de véhicule dont des **Volkswagen** ainsi que des SUV RX450h de **Lexus**. Début 2021, on apprenait qu'Apple pourrait lancer des véhicules électriques construits aux USA par **Hyundai**.



Autres services et fonctionnalités

L'IA peut aussi apporter des services en sortant du cadre des véhicules autonomes. En voici quelques exemples. Elle peut aussi servir à optimiser les trajets, notamment pour les professionnels. C'est un des domaines où **Uber** travaille, pour optimiser le temps de travail de ses conducteurs. **IBM** propose une solution « Watson on Wheels » qui optimise aussi les trajets en fonction d'informations sur la qualité de la voirie.

Une analyse des données issues des smartphones, de l'enregistreur de bord, de caméras, de la vitesse et du régime moteur permet d'évaluer le comportement du conducteur et, éventuellement, de moduler ses primes d'assurance en conséquence.

L'analyse du bruit du moteur permet de faire de la maintenance préventive. C'est une fonction proposée par la startup **Otosense** (2014, USA, $1,2M), qui est basée à Cambridge (USA) et créée par le Français Sébastien Christian[1060], chez **Carfit** (2016, USA, $2,3M), également créé par un Français, Henri-Nicolas Olivier (ex Modelabs et Inventel), et qui exploite l'analyse des vibrations du véhicule (schéma *ci-dessus*) ainsi que chez **Uptake Technologies** (2014, USA) qui l'a notamment expérimenté dans l'armée de terre US sur des chars Bradley M2A3[1061].

En Corée du Sud, **Hyundai** et **Kia** travaillent sur une boîte de vitesse automatique pilotée par du machine learning. L'idée est de minimiser le nombre de changements de vitesse et de réduire la consommation d'énergie en s'adaptant aux circonstances de la conduite[1062].

Geoptis (France) équipe les voitures des facteurs de La Poste avec des caméras qui servent à inspecter l'état du réseau routier pour détecter les nids de poule et les fissures[1063]. Reste ensuite à les réparer ! C'est une spin-off de La Poste qui vend cette solutions aux collectivités locales. La solution est complétée par un système d'optimisation du trajet des véhicules.

Stanley Robotics (France) développe un robot transporteur de véhicules en tout genre et notamment tourisme qui permet de les ranger automatiquement dans des parkings.

C'est une piste intéressante pour équiper les grands parkings, notamment dans les aéroports. Mais cela deviendra caduc lors de l'avènement des véhicules autonomes.

Knightscope (USA) est un concurrent américain indirect de Stanley Robotics qui propose un robot parking, en test sur un parking de Microsoft aux USA.

L'IA peut aussi servir à fluidifier le trafic dans une ville. Ainsi, la startup **Waycare** (2016, Israël/USA, $9,6M) a-t-elle créé une solution d'optimisation de gestion du trafic routier exploitant des capteurs embarqués dans les véhicules et sur les routes, caméras de surveillance comprises pour l'Interstate 15 à l'Ouest du Strip de la ville de Las Vegas[1064]. La méthode n'est pas précisée.

Accidents et responsabilités

Le premier accident mortel d'un véhicule équipé d'une fonction de conduite assistée est intervenu mi 2016 en Floride. Un procès s'en est suivi qui a dédouané Tesla. Le conducteur n'avait pas respecté les consignes de sécurité et les alertes[1065]. Mais le camions blanc dans lequel la Tesla s'était encastré n'était pas facile à éviter pour le capteur Mobileye de la Tesla. Le constructeur a donc fait évoluer sa configuration en multipliant les capteurs, passant notamment de un à huit capteurs RGB. L'absence de LiDAR pourrait cependant rester un handicap dans ces véhicules.

[1060] La startup proposait à l'origine un système transformant les alertes sonores en alertes visuelles pour les malentendants. C'est ensuite qu'elle s'est mise à cibler le marché industriel. Ils sont en test chez PSA. Leur projet AudioHound est une tablette conçue pour le technicien du garage pour réaliser des diagnostics pas analyse du bruit du véhicule.

[1061] Uptake Technologies est moins spécialisée que Carfit et Otosense. Ils ciblent tout un tas de marchés : l'agriculture, le rail, la construction, l'énergie et l'industrie.

[1062] Voir Hyundai and Kia Develop World's First ICT Connected Shift System, février 2020.

[1063] Voir Audit des routes avec les voitures de la Poste par Jacques Cheminat, novembre 2018.

[1064] Voir Artificial intelligence improves highway safety in Las Vegas, novembre 2018.

[1065] Voir The driver who died in a Tesla crash using Autopilot ignored at least 7 safety warnings de Brian Fung, juin 2017.

D'autres accidents mortels impliquant des véhicules autonomes ont eu lieu, améliorant la courbe d'expérience des industriels. Il y a eu coup sur coup un crash de Tesla en Chine en 2016 (vidéo), un véhicule autonome **Uber** renversant une dame marchant avec son vélo en Arizona en pleine nuit en mars 2018 (vidéo) et une **Tesla** Model X également le 23 mars 2018 en Californie[1066]. À ce jour, il y a donc eu au moins quatre accidents mortels impliquant des véhicules semi-autonomes[1067].

Pour l'accident impliquant Tesla et ayant entraîné le décès de son conducteur travaillant chez Apple, le NSTB a remis son rapport début 2020[1068]. S'il émet des recommandations concernant l'Autopilot et la gestion de la voie rapide concernée en Californie, le rapport avait tendance à faire porter le chapeau au conducteur qui se fiait trop à l'Autopilot et était probablement distrait par un jeu sur son smartphone.

Dans le cas Uber, le système de conduite autonome avait décidé d'ignorer l'objet qui avait bien été détecté devant le véhicule par les capteurs embarqués. Le système pensait que l'objet détecté ne pouvait pas être un piéton, ceux-ci traversant normalement sur les passages piétons ! Le conducteur du véhicule était distrait mais il n'est pas évident qu'il aurait évité la dame avec son vélo s'il avait été au volant, vues les conditions. Uber a annoncé avoir corrigé ce défaut dans le système de prise de décision de ses véhicules autonomes. En 2019, le NSTB remettait son rapport et mettait plutôt en cause la conductrice de sécurité qui accompagnait le chauffeur de ce Uber[1069]. En tout cas, même si ce n'est pas forcément directement lié à cette affaire Uber annonçait fin 2020 abandonner ses efforts dans la conduite autonome[1070].

Dans les autres accidents, les conducteurs n'auraient pas respecté les avertissements du véhicule.

Des accidents non mortels ont aussi été constatés. Avec par exemple un minibus du Français **Navya** percuté par un camion n'ayant pas respecté une priorité dans la zone de Fremont Street à Las Vegas en novembre 2017 ou une autre **Tesla** emboutissant un véhicule de police en Californie en mai 2018 et encore une Tela rentrant dans un camion de pompier dans l'Utah en mai 2018. A chaque accident, les commentateurs remettent en question la sécurité des véhicules autonomes et des procès sont déclenchés. Or, ce qui compte n'est pas d'avoir des accidents, mais d'en avoir moins au kilomètre parcouru qu'avec des véhicules classiques et de déclencher une courbe d'expérience permettant de faire baisser ce taux d'accident au km.

[1066] Voir Tesla Driver Died Using Autopilot, With Hands Off Steering Wheel de Dana Hull et Tim Smith, mars 2018.

[1067] C'est dans Wikipedia : https://en.wikipedia.org/wiki/List_of_self-driving_car_fatalities.

[1068] Voir Le mauvais usage de l'Autopilot Tesla épinglé par le NTSB par Philippe Schwoerer, février 2020 et Accident mortel en Tesla avec pilotage automatique : les responsabilités enfin établies par Eric Dupin, février 2020.

[1069] Voir Une erreur humaine est à l'origine de l'accident mortel causé par le véhicule autonome d'Uber par Valentin Cimino, novembre 2019 et le résumé du rapport du NSTB sur l'accident Uber en Arizona de 2018 : Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, NSTB, 2019 (5 pages).

[1070] Voir Uber abandons effort to develop personal self-driving automobile, décembre 2020.

Pour l'instant, il semblerait que les véhicules autonomes soient sûrs de ce point de vue-là. Il faut toujours adopter une vue statistique de la sécurité. On devra cependant tenir compte du fait que l'on tolère moins les accidents générés par des machines que ceux qui le sont par des erreurs d'origine humaine. On peut aussi réduire le risque d'accidents en utilisant du deep learning basé sur de l'apprentissage par renforcement et l'injection de données de stress de conducteurs humains[1071].

Des accidents impliquant des Tesla en mode Autopilot ont lieu maintenant à peu près tous les six mois. Ils sont souvent liés à des circonstances inopinées comme un véhicule à l'arrêt sur la voie gauche d'une autoroute. Mais on ne compare que rarement cela avec ce qu'aurait fait un conducteur humain[1072]. Une étude réalisée entre 2017 et 2018 montrait toutefois que ce genre de solution servait surtout à accompagner le conducteur et pas à le remplacer. L'autonomie n'est pas encore prête à gérer un grand nombre de cas de figure[1073].

Un grand débat a cours au sujet de la gestion des dilemmes par les véhicules autonomes en cas d'accident, lorsqu'il leur faudra choisir entre la mort certaine du conducteur, de ses passagers et de personnes sur la route ou entre plusieurs personnes différentes sur la route (enfants, adultes, femmes, bébés, seniors, prisonniers).

Une autre variation concerne l'autonomie des véhicules. Dans la vision « côte Ouest » des USA, les véhicules sont destinés à être totalement autonomes. Dans une vision « côte Est » ou européenne, on envisage plutôt une collaboration entre véhicules et infrastructures. Cette approche est poussée dans les concepts de smart city et notamment par les opérateurs télécoms et ceux du BTP.

Ce sont en fait des expériences de pensée assez éloignées de la réalité. Bien rares sont les conducteurs humains qui ont eu à gérer de tels dilemmes[1074]. Sachant par ailleurs que la gestion de ces dilemmes dépend de facteurs culturels qui ne sont pas homogènes dans le monde[1075].

Cela conduit cependant des chercheurs à proposer l'intégration de formes d'éthique dans les algorithmes et règles de fonctionnement des systèmes à conduite autonome[1076].

Impact macro de l'autonomie

Au-delà de ces questions théoriques d'éthique et même s'ils se déploieront par étape, et à une échéance encore incertaine, les véhicules autonomes produiront des transformations radicales de l'industrie automobile et de nombreuses industries adjacentes[1077].

Tout d'abord, il est fort probable que les véhicules personnels perdront de leur attrait pour une bonne part de la population, notamment en ville. Les flottes de véhicules autonomes avec une forte densité de circulation répondront plus vite à la demande en termes de temps de réponse qu'un véhicule personnel garé dans un parking qu'il faut aller chercher. Cela pourra remettre en cause la structure du métier de constructeur automobile.

[1071] Voir Visceral Machines: Risk-Aversion in Reinforcement Learning with Intrinsic Physiological Rewards par Daniel McDuff et Ashish Kapoor de Microsoft Research, septembre 2018 (11 pages).

[1072] Voir Tesla On Autopilot Slams Into Stalled Car On Highway, Expect More Of This par Lance Eliot, 2019 et Tesla's Latest Autopilot Death Looks Just Like a Prior Crash par Dania Maxwell, mai 2019.

[1073] Voir Electronic driving systems don't always work, tests show par Tom Krisher, novembre 2019.

[1074] Voir Robot Cars And Fake Ethical Dilemmas de Patrick Lin, Forbes, avril 2017. Qui explique que les dilemmes éthiques évoqués ne sont que des expériences de pensée théoriques qui présentent l'intérêt de pousser la réflexion aussi loin que possible.

[1075] Voir Should a self-driving car kill the baby or the grandma? Depends on where you're from. par Karen Hao, octobre 2018.

[1076] L'approche peut consister à générer un vote social pour identifier la préférence à intégrer dans l'IA. Est-ce de l'éthique pour autant ? Pas évident ! La foule est-elle toujours intelligente ? Vous avez deux heures ! Voir A Voting-Based System for Ethical Decision Making, septembre 2017.

[1077] Voir l'excellent article 73 Mind-Blowing Implications of Driverless Cars and Trucks de Geoff Nesnow qui inventorie 73 transformations liées au déploiement des véhicules autonomes.

Quelques autres exemples :

- Il y aura **moins de voitures en circulation** et moins d'embouteillages dans les villes dominées par les véhicules autonomes, notamment parce que les trajets des véhicules pourront être coordonnés de manière globale et centralisée. Une expérience réalisée à l'Université de Cambridge sur de petites voitures miniatures montrait que des véhicules qui collaborent génèrent moins de bouchons que ceux qui ne collaborent pas. Ce n'est pas une grosse surprise mais c'est une bonne nouvelle[1078].

- Il y aura, si tout va bien, **beaucoup moins d'accidents** [1079], avec un impact sur le marché des assurances d'un côté et aussi, sur les systèmes de santé qui seront en partie désengorgés.

- L'usage de véhicules autonomes sera accompagné d'une migration à la **propulsion électrique**, avec un impact positif sur la qualité de l'air dans les villes[1080].

- Le **tourisme** pourrait être affecté, rendant certains sites plus facilement accessibles, notamment aux personnes à mobilité réduite. Des véhicules autonomes pourraient aussi prendre en charge des parcours à la carte avec des départs n'importe où et une durée arbitraire fixée par les touristes. Certains imaginent même que les véhicules autonomes pourraient servir de maisons closes ambulantes, notamment dans des villes telles qu'Amsterdam[1081].

- Les **parkings** pourront être plus compacts, ceux-ci ne nécessitant pas d'être accessibles par les passagers.

- Les passagers pourront vaquer à d'**autres occupations** dans leur véhicule, qu'il s'agisse de travail ou de loisir. Les techniques de personnalisation numérique de l'environnement des véhicules se développeront.

- Cette migration se fera au prix d'un changement des **méthodes de production d'énergie**, demandant plus de production d'énergie électrique, avec des énergies intermédiaires stockables, ce que ne permettent pas forcément les énergies renouvelables issues du soleil (photovoltaïque et éolien). Des **métiers disparaîtront** comme avec la fin des diligences et du cheval de transport ! Les métiers de la conduite représenteraient plus de 4 millions d'emploi aux USA[1082], soit de 1,6% à 3,5% des emplois selon les états. Les radars disparaîtront aussi et on n'aura plus besoin de **permis de conduire**[1083] !

[1078] Voir +35 à +45% de fluidité sur routes avec des véhicules autonomes coopératifs par Korben, 2019 et A Fleet of Miniature Cars for Experiments in Cooperative Driving par Nicholas Hyldmar, Yijun He et Amanda Prorok, février 2019 (7 pages).

[1079] La route fait environ 1,25 millions de morts par an dans le monde, concentrés en Inde et en Chine en volume (260 000 pour cette dernière), puis aux USA ! Plus que n'importe quelle guerre. Il faut y ajouter entre 20 et 50 millions de blessés par an. Voir Road Safety Facts.

[1080] Voir Using ICs To Shrink Auto's Carbon Footprint par by Ann Steffora Mutschler, décembre 2020 qui porte sur l'impératif de baisse de consommation de l'électronique embarquée dans les véhicules mais aussi sur l'impact de ces composants sur les économies d'énergie.

[1081] Voir Self-driving cars could be the brothels of the future, UK researchers say - Roadshow par Kyle Hyatt, novembre 2018 et Autonomous vehicles and the future of urban tourism par Scott Cohen, 2019.

[1082] Dans Stick Shift : Autonomous Vehicles, Driving Jobs, and the Future of Work, 2017 (40 pages).

[1083] La liste de cette page provient de 128 Things that will disappear in the driverless car era, avril 2016.

Driving Jobs	Agriculture	Miscellaneous Jobs	Justice	Highway Related
Taxi drivers	Tractor drivers	Traffic reporters on the news	Traffic cops	Traffic jams
Uber & Lyft drivers	Combine operators	Sobriety checkpoint people	Traffic courts – lawyers, DA, judges	Traffic signs
Delivery (FedEx, UPS, USPS) jobs	Swather operators	Auto industry lobbyists	Driver licenses	Traffic lanes
Courier jobs	Bailer operators	Stoplight installers	Patrol cars and officers	Speed zones
Bus drivers	Sprayer operators	Pothole repair people	DUIs and drunk driving	Road stripes
Truck drivers	Horse trailer drivers	Emission testers	Sobriety checkpoints	Weigh stations
Valet jobs	Grain truck operators	Road and parking lot stripers	The boot	Mile markers
Chauffeurs and limo drivers	Automated fruit harvester operators	Night repair crews	Road rage school	Guardrails
Other Jobs	**Construction**	**Vehicle Repair**	**Crash test dummies**	**Highway Repair**
Road construction flag people	Crane operators	Roadside assistance	Road rage	Traffic cones
Drivers-Ed teachers	Road grader operators	Auto repair shops	Fender benders	Road closures
Defensive driving schools	Earth movers	Body shops	Car theft	Detours
Traffic analysts	Street sweeper operators	Tow trucks	Getting lost	Stoplights
Car licensing and registration	Backhoe operators	Glass repair	Lost cars in parking lots	Pilot cars
Drivers test people	Trencher operators	Auto locksmiths	Driving tests	Flag people
Rental car agents	Cement truck operators	Transmission repair shops	Traffic stops	Merge lanes
Crash testers	Fuel truck operators	Auto part stores	Crash test dummies	Night lights for late night road repair
Specialty Vehicles	**Car ecosystem**	**Vehicle Maintenance**	**Parking Related**	**Traffic Laws**
Forklift drivers	Auto sales – new and used	Gas stations	Parking lots	Speeding tickets
Lawnmower operators	Account managers	Car washes	Parking garages	Failing to stop at a stoplight or stop sign
Snowplow operators	Auto auctions	Oil change businesses	Parking tickets	DUIs – driving under the influence
Water truck drivers	Credit managers	Detail shops	Valet services	Reckless driving
Fire truck drivers	Loan underwriters	Tire shops	Parallel parking	Driving in the wrong direction
Water taxies	Insurance agents and sales reps	Brake shops	Parking meters	Passing in a no passing zone
Ambulance drivers	Insurance claims adjusters	Emissions testing	Charging stations	Unsafe lane changes
Trash truck drivers	Insurance call center agents	Alignment shops	Handicap parking	Driver profiling

- Il y aura moins d'accidents mortels sur la route, ce qui aura un impact indirect sur la source **d'organes pour les greffes**[1084] issus de donneurs en bonne santé. Un beau paradoxe à gérer.
- Les **villes** pourront être réaménagées. Les temps de trajets seront davantage prédictibles et l'intermodalité plus facile à mettre en œuvre. Cela rendra la vie des banlieusards plus acceptable et aura un impact sur le marché immobilier. On aura par contre encore besoin de feux de circulation pour laisser passer les piétons !
- Cela deviendra un éminemment **sujet politique**. Il provoquera des résistances et des débats qui rappelleront ceux qui sévissent autour de la 5G. Certains voudront même interdire les véhicules autonomes, déclencher des moratoires au premier accident venu, déployer à outrance le principe de précaution. D'autres s'inquiéteront de l'impact énergétique global de l'autonomie qui ne serait pas si rose que cela[1085]. Les industriels devront s'y préparer !

Rail

Comme pour le transport routier, le rail peut faire appel à de l'IA pour son autonomisation mais d'autres usages voient aussi le jour, notamment autour de la maintenance des infrastructures et du matériel roulant.

Autonomisation

Le rail utilise aussi des véhicules autonomes et depuis longtemps, ne serait-ce que dans les métros des aéroports, OrlyVal ou les lignes du métro 1 et 14 à Paris, et bientôt 4. La sécurité est assurée par les doubles portes de protection pour accéder aux wagons. Les utilisateurs savent aussi généralement s'adapter aux automatismes et jouent moins avec les portes pour rentrer au dernier moment dans les wagons.

L'automatisation du rail pourra un jour également toucher les lignes de train. A ce titre, la **SNCF,** l'IRT **Railenium**, **Alstom**, **Altran**, **Ansaldo**, **Apsys**, **Bombardier**, **Bosch**, **Spirops** et **Thales** (ça fait du monde !) annonçaient en septembre 2018 la création de deux consortiums visant à développer d'ici 2023 deux démonstrateurs de trains autonomes, un TER et pour le transport de fret. On en est pour l'instant au stade des avant-projets. Cela entraînera entre autres une refonte des systèmes de signalisation, de contrôle-commande et l'optimisation de l'exploitation ferroviaire.

[1084] Voir Plus de voitures autonomes, c'est aussi moins d'organes pour les greffes d'Ian Adams et Anne Hobson, novembre 2018.

[1085] Voir Lettre aux ingénieurs du véhicule autonome : « Je vous écris parce que c'est de nos vies à tous qu'il s'agit » par Celia Izoard, octobre 2020.

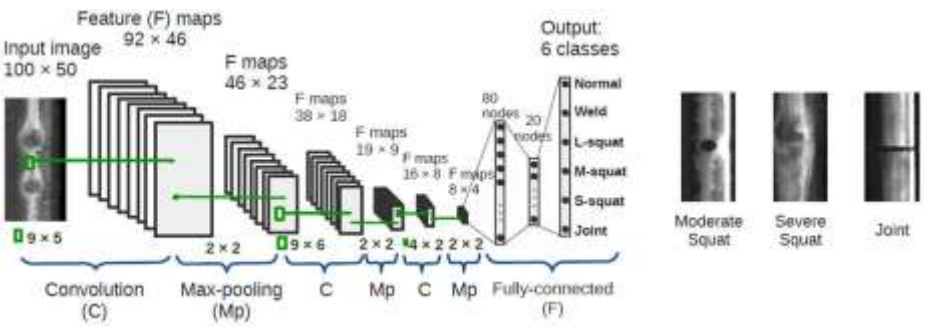
Maintenance et opérations

Sinon, le monde du rail est depuis longtemps un utilisateur de l'IA. Il a commencé à expérimenter les systèmes experts dans les années 1980. Il est surtout focalisé sur l'optimisation des ressources et la maintenance prédictive.

Les systèmes experts et la logique floue peuvent servir à la planification, à l'évaluation des retards[1086] et à la reprogrammation de trains après des retards. D'une manière générale, les transports nécessitent des outils de plus en plus sophistiqués pour gérer le matériel et les Hommes qui les opèrent. Toutes les données issues des nombreux capteurs sur le matériel roulant et dans les infrastructures peuvent permettre de faire de la maintenance prédictive et l'améliorer la disponibilité des infrastructures ferroviaires. Le déploiement de ces solutions est cependant évidemment contraint par les ressources financières des opérateurs ferroviaires.

J'ai trouvé dès 2013 des traces d'inspection visuelle des voies ferrées à base de réseaux de neurones, ce qui correspond au début de l'engouement pour la technique[1087]. Elle peut porter sur l'inspection des rails, du ballast et des environs. L'objectif est d'identifier des obstacles sur les voies, les défauts de structures sur les rails et la croissance de la végétation[1088]. L'inspection peut être réalisée par des caméras installées sur des motrices dédiées à la maintenance des voies et même sur des motrices tirant les trains habituels[1089].

Aux Pays-Bas, la détection de défauts sur des rails s'appuie sur des réseaux de neurones convolutifs exploitant des images à basse résolution (100x50 pixels) pour classifier une demi-douzaine de défauts répertoriés[1090].



La recherche de défauts sur les rails, les caténaires, dans la signalisation et dans le matériel roulant fait partie des domaines de prédilection de l'IA[1091]. Des accidents comme celui de Brétigny en 2013 ont accentué les efforts de la SNCF en France dans la surveillance des infrastructures. La SNCF inspecte les caténaires de ses lignes avec des drones et de la classification des images des isolateurs en céramique ou en verre.

[1086] Des projets de prédiction des retards et d'information voyageurs associés étaient expérimentés à la SNCF en 2019, s'appuyant sur du machine learning et des données d'entraînement. C'est un projet très complexe qui nécessite d'intégrer de nombreuses données disparates à structurer et nettoyer. Dans ce domaine-là aussi, les données d'entraînement peuvent générer des biais qu'il faut essayer d'éliminer.

[1087] Voir Automated Visual Inspection of Railroad Tracks de Esther Resendiz, John Hart et Narendra Ahuja, 2013 (10 pages).

[1088] Voir Automating Condition Monitoring of Vegetation on Railway Trackbeds and Embankments de Roger Nyberg, 2015 (301 pages).

[1089] Voir Digital Twin for the Railway Network Making Trains "Look" for Track Defects de Dattaraj Jagdish Rao, GE Transportation, 2018 (16 slides).

[1090] Source : Big Data in railway infrastructure, 2018 (52 slides).

[1091] L'illustration d'inspection de matériel roulant provient de la startup PointR, partenaire d'IBM (vidéo).

La maintenance prédictive peut aussi exploiter de nombreuses données issues de capteurs intégrés dans le matériel roulant[1092].

C'est ce que fait la **SNCF**, avec **Quantmetry** (2010, France), **Global Sensing Technologies** (2011, France) et **IBM Watson** pour la maintenance des Transiliens depuis un pilote lancé fin 2015.

Enfin, la relation client grand public passe évidemment par les chatbots, même si ceux-ci sont encore loin d'être au point comme l'illustre l'exemple ci-contre avec celui de **Ouisncf** qui ne sait pas gérer une demande sur un aller/retour et ne mémorise pas la date de la première demande[1093] !

Offres clients

Du côté de l'offre utilisateurs, la notion d'intermodalité entraîne des évolutions de l'offre associant différents opérateurs de transports. On en profite déjà largement via les applications transverses mobiles, à commencer par **Google Maps** et autres **CityMapper**.

Geo4cast (2012, France) propose dans ce domaine des services d'optimisation des transports pour tous les modes. La startup récupère des données de déplacements issues de smartphones, est connecté avec des données automobiles, transports en commun et permet d'analyser et d'optimiser les transports entre eux pour réduire les investissements et dépenses superflues. Cela pourrait servir à terme la coordination automatique de réseaux de véhicules autonomes. Le tout exploite plusieurs technologies comme le nettoyage de données, la reconnaissance de modes de transports basée sur du machine learning, des algorithmes d'optimisation et du « revenue management ».

Aviation

Le monde du transport aérien est aussi un grand consommateur d'IA en pièces détachées, que ce soit pour le *yield management* appelé aussi *revenue management* qui exploite un mix de *big data* et d'IA aussi bien dans les compagnies aériennes que chez Amadeus et son service de réservation en marque blanche, la gestion des opérations et la relation client[1094].

Il en va de même pour les avions eux-mêmes qui font appel à divers systèmes embarqués qui comprennent très probablement de nombreuses briques d'IA, ne serait-ce que pour la navigation. Leur maintenance prédictive peut aussi faire appel à de l'IA.

Enfin, les aéroports qui sont de véritables villes avec des dizaines de corps de métiers et services à coordonner sont également de grands consommateurs d'IA.

[1092] Voir Using IOT to Advance Railway Predictive Maintenance de Hitachi Vantara, 2018 (28 pages) et IoT and predictive maintenance in Railways, Thalès, 2018 (14 slides).

[1093] Ce qui n'a pas empêché ce chatbot de gagner le prix du « Best Robot Experience » en avril 2018. Voir aussi Comment Oui.sncf propulse son bot sur Google Home, Alexa ou Messenger d'Antoine Crochet-Damais, mai 2018.

[1094] Voir Impact of AI on the Aviation Industry de Houman Goudarzi, 2017.

La création de solutions d'IA peut s'appuyer sur les tombereaux de données générées par les opérations, aussi bien au niveau des capteurs intégrés dans les avions, les données météo et d'opérations qu'au niveau des passagers transportés ou dans les aéroports[1095]. Un patchwork complexe de règles métiers complète le tout et peut être intégré dans des briques d'IA à base de moteurs de règles.

Comme dans tous les autres marchés, le secteur aérien exploite ou exploitera des solutions d'IA, qui sont soit isolées et relativement indépendantes comme un système de vidéosurveillance, et d'autres qui relèvent de l'intégration de données techniques disparates, comme pour prévoir les ressources à aligner dans les aéroports pour accueillir les avions et leurs passagers. La majorité des solutions d'IA évoquées ici sont des outils d'analyse et d'aide à la décision et au pilotage des opérations. Rares sont celles qui sont entièrement autonomes.

Le marché de l'IA dans le secteur de l'aviation passerait de $112,3M en 2017 à $2,2B en 2025, avec une croissance annuelle de 46,65% probablement équivalente à celle d'autres secteur d'activité[1096]. Sachant que le revenu des compagnies aériennes était situé aux alentours de $830M en 2018. C'est une secteur économique très concurrentiel, très réglementé et avec des marges tirées au cordeau.

Compagnies aériennes

Les compagnies aériennes génèrent un business récurrent qui doit tourner sans failles et assurer la sécurité et le confort des passagers. On peut segmenter les processus et usages de l'IA en trois grandes catégories : la vente, l'exploitation et le service au client.

La vente de billets d'avions s'appuie depuis des décennies sur des techniques d'optimisation de la tarification, le fameux **yield management**. Celui-ci fait appel à d'innombrables méthodes mathématiques qui ont petit à petit intégré celles du machine learning. Cela relève de la programmation par contraintes dans un environnement instable : comment maximiser le revenu et le taux de remplissage des avions en fonction des destinations, de l'évolution de la demande, de la concurrence, de la météo et d'autres paramètres divers. Le machine learning permet notamment d'identifier des corrélations entre ces différents paramètres et le résultat escompté. Le yield management peut aussi intégrer des systèmes de recommandation de vols pour les *frequent fliers*.

Yieldin (2013, France) fait du « revenue management » à base d'IA, non précisée, probablement avec une grosse dose de machine learning et surtout avec une interface utilisateur métier adaptée aux tâches d'optimisation des revenus de l'usage des capacités de vol. Leurs clients sont des compagnies aériennes de « tier 2 » telles que XL Airways (qui a fait faillite en 2019), Air Malta et Monarch Airlines (UK).

Lors de la vente de billets en ligne, des systèmes de **détection de fraudes** diverses peuvent être mis en jeu, comme dans le commerce en ligne mais en tenant compte de la spécificité de la vente de billets d'avions. Elle doit notamment intégrer les pratiques de certains professionnels qui peuvent acheter des billets en volume à la baisse et les revendre à la hausse.

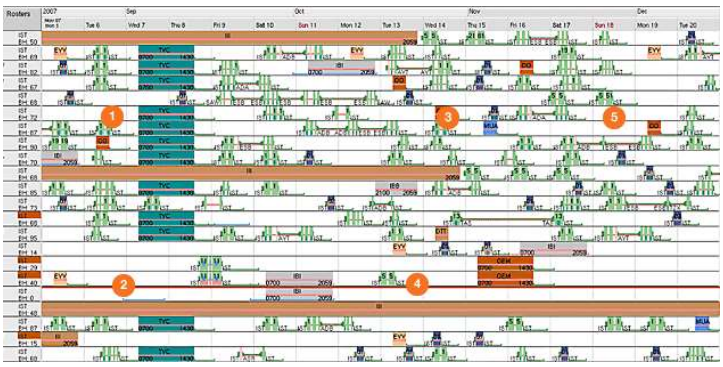
[1095] Voir Machine Learning in Aviation, 2013 (54 slides).

[1096] La source est Artificial Intelligence in Aviation Market by Offering qui est citée dans la présentation Artificial Intelligence and Machine Learning in Aviation 2018 (337 slides).

Suivent les applications de **gestion des opérations**. Elles comprennent les outils de prévision des retards servant ensuite à adapter et replanifier automatiquement les parcours de passagers et optimiser la gestion des aéroports « hubs » de la compagnie aérienne, comme Roissy CDG pour Air France ou Atlanta pour Delta.

L'IA peut aussi intervenir dans les algorithmes d'optimisation des routes prises par les avions pour optimiser les délais et réduire la consommation de kérosène tout en maximisant le confort des passagers en évitant les zones de turbulence. Cela relève encore de techniques associant de la programmation par contraintes et du machine learning pour l'apprentissage à partir de données historiques.

On peut optimiser la gestion du personnel navigant et prédire les retards des avions[1097]. Ainsi, **Jeppesen** (1934, USA), une filiale de Boeing, est à l'origine de Crew Rostering, une solution logicielle qui sert à gérer le planning du personnel navigant en intégrant différentes contraintes comme leurs préférences et leur niveau de fatigue. Elle utilise des outils de machine learning.



C'est sinon le spécialiste mondial des cartes de navigation aérienne pour les pilotes de ligne.

Un système d'analyse des paramètres de vol a été déployé à partir de 2011 par **Southwest Airlines** pour identifier les anomalies dans les rapports de vol des pilotes et les prévenir ainsi que les équipes de maintenance au sol, de potentiels problèmes. Le système a été développé à partir d'une solution de DMS (Data Management System) de la NASA exploitant des séries temporelles[1098]. Des analyses multi variables permettent de déterminer des relations de causalité complexes. Ainsi, ils ont découvert que la surconsommation de carburants sur certains vols provenait de défauts de capteurs sur un modèle particulier d'avion. À l'**Onera**, on exploite des techniques de machine learning à base d'arbres de décision pour analyser a posteriori les vols et leur sécurité au niveau de leur enveloppe de vol[1099].

La **relation client** passe de plus en plus par des chatbots et des agents vocaux. Amazon Alexa est utilisé chez United. **OpenJaw Technologies** (2002, Irlande) créé des outils de relation client pour les compagnies aériennes. T-Social est un chatbot développé avec les briques de traitement du langage d'IBM Watson. Il comprend un système d'escalade des cas non résolus vers des opérateurs humains. La société fait partie du groupe TravelSky Technology depuis 2016. **eDreams** propose un chatbot pour accéder aux actions transactionnelles liées aux réservations. J'ai pu le tester pour récupérer une facture d'un vol à l'étranger.

Air France a mis en place divers **chatbots** pour la relation client passagers et l'application Cargo Repair Case pour la réallocation optimisée de charges cargo dans les avions de passagers en fonction des contraintes et aléas d'exploitation.

Chez Air France, l'IA est aussi testée pour aider les passagers à trouver les zones d'enregistrement et de livraison des bagages dans les aéroports avec une application de réalité augmentée pour mobile.

[1097] Voir Application of Machine Learning Algorithms to Predict Flight Arrival Delays, 2017 (6 pages).

[1098] Voir Data Mining Tools Make Flights Safer, More Efficient, 2013.

[1099] Voir Modeling airline crew activity to improve flight safety analysis de Nicolas Maille, Onera, 2017 (12 pages).

EasyJet optimise de son côté la vente de produits à bord des avions, surtout les boissons et repas. Ils exploitent des techniques de machine learning de prévision de la quantité de produits à embarquer en fonction de divers paramètres comme la météo et les caractéristiques du vol (de jour, de nuit).

Enfin, le machine learning et le traitement du langage permet l'analyse de l'opinion dans les réseaux sociaux, comme pour toutes les entreprises ayant pignon sur rue. **PureStrategy** (2015, Canada) est un fournisseur d'outil d'analyse de données à base de machine learning, l'Automated Neural Intelligence Engine qui peut servir à collecter et analyser le feedback des passagers sur leur expérience clients. C'est une startup de moins de 10 personnes !

Aéroports

Les aéroports ne sont pas en reste pour les usages de l'IA. Ce sont des plaques tournantes au propre et au figuré.

Ils hébergent en effet un grand nombre d'acteurs à coordonner entre eux : les compagnies aériennes, les transports terrestres (train, taxis), les services de police et de douane, le contrôle de sécurité à l'embarquement, les prestataires divers sur les pistes, le catering des avions, les commerçants et restaurateurs, etc.

La première application courante de l'IA dans les aéroports concerne les usages de la **reconnaissance des personnes et des visages.** Ils sont maintenant couramment intégrés dans les bornes de l'immigration aux USA, comme en France avec Paraphe ou à Heathrow. Les aéroports des Emirats Arabes Unis prévoient d'aller jusqu'à supprimer les agents de contrôle des passeports et de se reposer sur des systèmes entièrement automatiques, à base de caméras et de reconnaissance de visages[1100].

La reconnaissance de visage commence aussi à faire son apparition pour le contrôle des passagers à l'embarquement en lieu et place des cartes d'embarquement[1101].

À l'aéroport d'Atlanta depuis fin 2018, elle est utilisée pour la dépose des bagages aux comptoirs de Delta Airlines, pour le contrôle de sécurité (TSA) et pour l'embarquement. Le tout grâce à votre photo enregistrée dans la base du CBP (Customs & Border Patrol).

L'IA et la reconnaissance faciale peuvent être combinées aux scanners de contrôle de sécurité des passagers comme ceux d'**Evolv Technology** (2013, USA, $42,4M) qui détectent bombes et explosifs grâce à des ondes millimétriques.

CrowdVision (2009, UK) aide aussi les aéroports à gérer leurs flux de passagers, à optimiser les temps de check-in et de passages aux points de sécurité. La solution exploite la vidéo surveillance et l'analyse des flux de passagers dans l'aéroport. A la clé, la réduction des files d'attente, l'orientation des passagers et l'amélioration de la sécurité (vidéo). Les réseaux multi-agents permettent de leur côté de simuler le comportement de foules dans des aéroports en cas d'incidents ou d'événements anormaux. Ils peuvent servir à améliorer la conception des aéroports.

L'analyse des habitudes du personnel au sol, associée à des solutions de machine learning, peut servir à détecter des anomalies comportementales ayant une incidence sur la sécurité des vols[1102].

[1100] Voir Immigration officers in the UAE will be replaced with AI by 2020, février 2019.

[1101] Voir How can AI help speed up airport security?, janvier 2019.

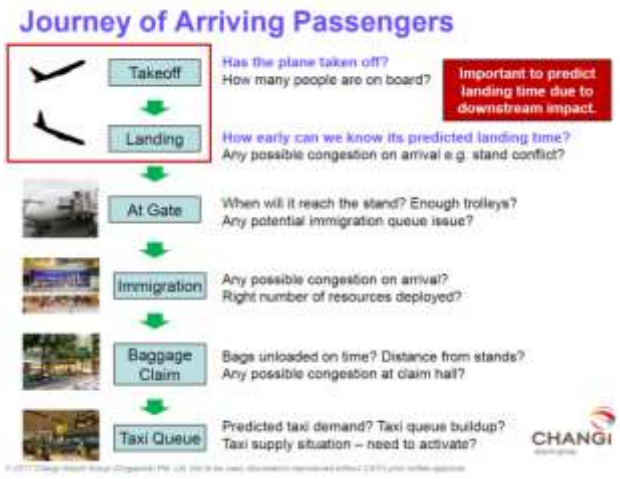
L'IA peut aussi jouer un rôle dans le **contrôle au sol de l'aéroport**. Le canadien **Searidge Technologies** (2008) propose IntelliDAR, une solution de contrôle au sol qui exploite l'image de caméras et du stitching vidéo (vidéo). Elles permettent la détection et le positionnement de tous les engins mobiles sur les pistes et un contrôle au sol à distance.

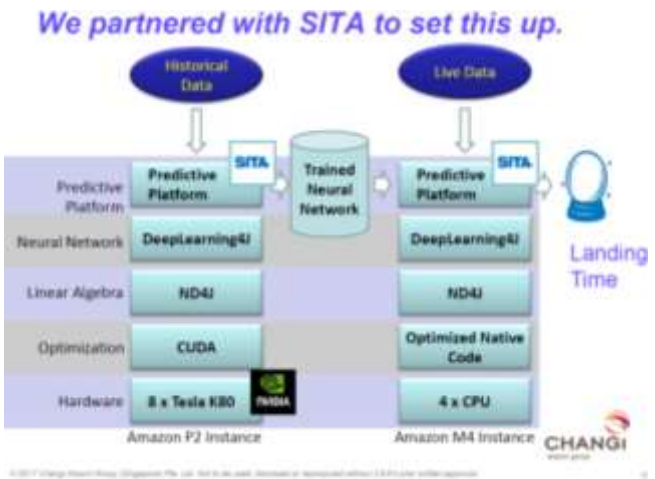
En complément, leur solution **Aimee** gère le contrôle aérien aéroportuaire. Ses briques d'IA servent à détecter, classifier et suivre les avions, à analyser les dialogues pilotes/tour de contrôle, à surveiller les plans de vols et à intégrer les contraintes issues de la météo[1103]. Le réseau de neurones de reconnaissance d'avions est entraîné avec des jeux de données 3D synthétiques[1104]. Le tout exploite de la réalité augmentée pour la visualisation.

Le contrôle aérien devrait de son côté pouvoir exploiter des outils de reconnaissance de la parole capables de tenir compte de la piètre qualité des communications et d'identifier des erreurs d'interprétation éventuelles des instructions du contrôle ou des équipages.

Le projet et la joint-venture **Skygrid** lancé par Boeing et l'éditeur de logiciels **SparkCognition** vise à créer une plateforme logicielle de routage automatique et dynamique des avions civils et cargo en s'appuyant sur du machine learning et sur une blockchain dédiée.

Les aéroports doivent investir aussi pour prédire la manière dont leurs infrastructures seront sollicitées, afin de faire du « capacity management ». D'où par exemple, le besoin de prédire l'heure d'arrivée des avions qui dépend d'un grand nombre de paramètres et va conditionner l'usage des services de bagage, de contrôle des passeports et les transports terrestres[1105].





[1102] Voir Airports turn to Artificial Intelligence to find the dangers within, 2018.

[1103] La société documente bien son usage de l'IA dans Artificial Intelligence in ATM de Marco Rueckert, Searidge Technologies, 2017 (18 pages). Les techniques d'IA employées pour la gestion des pistes relevaient au depart de moteurs de règles, comme évoquées dans An Application of Artificial Intelligence for the Safety in the Neighbourhood of Airport Runways par Christophe Blaess (consultant) et Claude Tsiampalidis (ADP), 1996 (6 pages). On complète cela depuis avec de la vision artificielle. On le constate également dans Artificial intelligence with applications for aircraft, de la FAA, 1994 (186 pages).

[1104] L'IA est utilisée depuis longtemps dans le contrôle aérien. Dans les années 1980, la technique utilisée était celle des moteurs de règles et systèmes experts. Voir Applying Artificial Intelligence Techniques' to Air Traffic Control Automation, 1989 (18 pages).

[1105] Voir AI in a SMART Airport par Steve Lee de l'aéroport de Changi (Singapore), 2017 (18 slides).

La **gestion des bagages** est l'une des principales tâches réalisées par les aéroports pour le compte des compagnies aériennes. L'objectif est évident : réduire les erreurs de routage des bagages et le temps d'attente des passagers puis gérer efficacement les exceptions. La vidéo-surveillance des aéroports permet aussi d'identifier l'origine des bagages abandonnés par les passagers et de faciliter l'intervention des services de déminage.

Fieldbox.ai (2011, France) propose une boîte à outil d'IA pour la gestion de processus industriels divers. Elle sert à collecter des données opérationnelles, à les présenter et à détecter des anomalies et tendances. Elle est évaluée par Aéroports de Paris pour analyser le processus de tri des bagages et l'optimiser (vidéo).

La gestion des bagages peut aussi par la case de la robotisation. Ainsi, le **BagBot** de DGWorld remplit automatiquement les containers de valises dans quelques aéroports européens depuis 2014, notamment à Heathrow (vidéo).

DigiBot BagBot (2014) Schiphol

Skywash (1997) Putzmeister

Divers robots de lavage des avions sont aussi en service. Le robot **Skywash** de Putzmeister lave les avions de toute taille depuis 1997, notamment à Frankfort en Allemagne. Il est construit autour d'un robot AEG. On en trouve aussi chez **Nordic Dino** d'**Aviator Robotics** (2010, Finlande) (vidéo). Le positionnement du Skywash exploite des lasers. Ces robots ont l'air d'être plutôt des automates que des engins autonomes.

Même si leur usage est soumis à une réglementation, des drones peuvent servir à inspecter les pistes et à détecter des objets intrus et à analyser la qualité du revêtement pour le réparer rapidement. La technique était expérimentée chez ADP à Roissy depuis 2016 avec la société **Airsight** (2012, Australie). Elle mobilisait les pistes pendant seulement 18 minutes[1106].

Les transports terrestres aboutissant et partant des aéroports peuvent aussi donner lieu à des applications de l'IA. Cela commence avec la reconnaissance visuelle des plaques d'immatriculation des véhicules entrant dans les parkings permettant notamment d'identifier des véhicules volés en sortie.

À moyen terme, on pourra voir apparaître des **véhicules autonomes** dans les aéroports : pour le transport des passagers, des bagages[1107] et des escaliers d'embarquement.

[1106] Voir Runway pavement inspections using airsight drone,Airsight, 2016.

[1107] Voir L'aéroport de Toulouse teste le tracteur bagages autonome de Charlatte et Navya par Léna Corot, décembre 2019.

En plus, bien entendu, de l'accueil des véhicules autonomes dans les terminaux de passagers et probablement de l'aménagement des voies d'accès pour ces véhicules. Sans compter l'éventuel avènement de drones de passagers qui apportera son lot de contraintes de coordination et notamment de contrôle aérien.

Les aéroports peuvent enfin améliorer le service aux passagers avec des chatbots, des applications de réalité augmentée pour mobile permettant de s'orienter dans l'aéroport (elle est proposée par Air France) voire avec des robots d'accueil comme c'est courant en Asie, qui peuvent par ailleurs faire le ménage lors de leurs déambulations[1108]. Des solutions de parking robotisées voient aussi le jour comme celle de **Stanley Robotics**, déjà citée un peu plus haut[1109].

La dimension commerciale de l'aéroport est particulièrement critique dans leur modèle économique. Diverses solutions à base d'IA peuvent ainsi servir à améliorer leur qualité et leur rendement, et qui relèvent plutôt de la catégorie « retail » physique, que nous évoquons à partir de la page 490.

Avions

Les avions eux-mêmes font appel à différentes formes d'IA. Par certains côtés, le bon vieux **pilote automatique** en fait partie, même s'il repose sur des moteurs de règles et des systèmes d'optimisation assez anciens. On y ajoute maintenant des briques de vision artificielle[1110].

Depuis 2016, la **DARPA** expérimente le concept du robot co-pilote, dans le cadre de son programme ALIAS (Aircrew Labour-in-Cockpit Automation System). Il contrôle les commandes de à la place du co-pilote. Les tests les plus récents avaient lieu sur des hélicoptères de Sikorsky dont un BlackHawk. L'idée est de créer une cohabitation entre le pilote et le robot co-pilote pour délester le premier des tâches basiques. Le pilote devient alors un chef de mission qui en contrôle l'exécution.

Dans l'aviation civile, Thales développe PureFlyt, un système de co-pilotage de l'avion utilisant du machine learning entraîné avec des logs d'une centaine de million d'heures de vol[1111].

La **maintenance des avions** peut aussi tirer parti de l'IA pour analyser et optimiser la performance des moteurs[1112] et pour inspecter visuellement les carlingues d'avions[1113] (*ci-contre*, un jeu d'entraînement de surfaces défectueuses à gauche et non défectueuses à droite). Air France fait de la maintenance prédictive de sa flotte de Boeing et d'Airbus au niveau moteur et avion qui permet de remplacer des pièces critiques avant leur défaillance[1114].

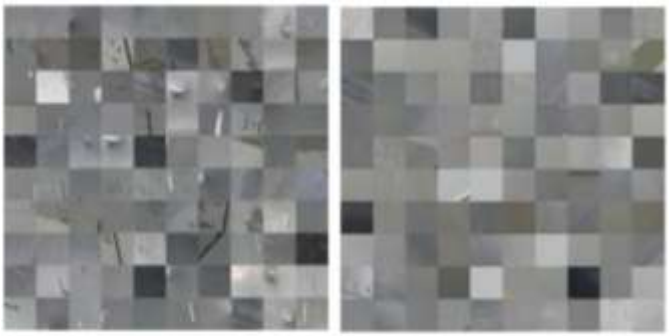


Figure 2. Examples of defect (left) and no-defect patches (right)

[1108] Voir par exemple le vidéo Robots assist passengers, clean floors at South Korean airport concernant l'aéroport de Séoul en 2017.

[1109] Voir Rhône. Aéroport Lyon Saint-Exupéry : le parking robotisé gagne du terrain, février 2020.

[1110] Voir Aircraft lands itself truly autonomously for the first time, juillet 2019 qui porte sur un test de solution d'atterrissage à base de systèmes de guidage (GPS) et de vision artificielle.

[1111] Voir Thales dévoile PureFlyt, le cerveau connecté de l'avion de demain, décembre 2019.

[1112] Voir le projet de Rolls-Royce dans Rolls-Royce To Use Artificial Intelligence For Predicting Engine Performance, mai 2018. Il est réalisé avec la startup **Uptake** (2014, USA, $218M) qui est spécialisée dans l'IA appliquée à l'industrie autour d'une plateforme maison de machine learning. L'application fait de la prévision de performance pour les moteurs Trent qui équipent les Boeing 787 Dreamliner et l'Airbus A330neo.

[1113] Voir Aircraft Fuselage Defect Detection Using Deep Neural Networks, 2017 (5 pages).

[1114] Voir PROGNOS - Predictive Aircraft Maintenance et la vidéo associée.

Airbus collecte les données d'exploitation des avions avec son programme Skywise, issues des plus de 30 000 capteurs des avions de ligne qui génèrent plusieurs To de données par journée d'exploitation. C'est une plateforme qui recueille les données de 6000 avions en service chez 80 compagnies aériennes en liaison avec l'outil d'analytics Palantir Foundry. Le programme a été lancé en 2017 et est notamment exploité par des partenaires services et intégration d'Airbus comme Accenture, Capgemini, FPT Software, IBM et Sopra Steria. Delta l'utilise.

SynapseMX (2015, USA) permet d'optimiser la maintenance des avions avec une application mobile destinée aux techniciens de maintenance. Ils utilisent du machine learning pour aider les techniciens « à prendre les bonnes décisions ». L'un de leurs clients est la compagnie aérienne low-cost AirTran basée à Atlanta.

Paladin AI (2013, Canada, $800K) crée des outils basés sur les données pour optimiser et accélérer la formation en aviation. Utilise l'apprentissage machine basé sur le nuage et l'IA pour apprendre dynamiquement les modèles de comportement indiquant la compétence du pilote.

Passagers

On peut aussi trouver quelques startups proposant des solutions à base d'IA s'adressant directement aux passagers sans passer par les compagnies aériennes :

- **Appintheair** (2011, USA, $300K) avec son application mobile d'intégration de l'expérience du voyageur aérien qui gère notamment les cartes d'embarquement, les horaires et la recherche de services dans les aéroports. Le tout est accessible par un chatbot pour Facebook Messenger.
- **Mezi** (2015, USA, $11,8M) avec son agent conversationnel qui aide les voyageurs à gérer … leur voyage *top to bottom*, et via SMS. La startup a été acquise par American Express en janvier 2018.

Mer

La mer et en particulier le transport maritime ne sont pas laissés pour compte de l'IA et, surtout, du machine learning.

Transport maritime

Le transport maritime s'intègre dans de complexes chaînes logistiques impliquant les industries manufacturières et les canaux de distribution et à l'échelle mondiale. Les volumes d'activité du transport commercial maritime et leur côté routinier permet de faire de nombreuses prévisions et de déployer diverses stratégies d'optimisation. C'est en particulier le cas du transport de containers qui peut être optimisé d'un bout à l'autre de la chaîne[1115].

Le machine learning peut servir à produire des prévisions plus fines des besoins, des délais de livraison[1116], à améliorer la sécurité dans les zones à fort trafic (c.f. l'exemple de détection d'anomalie dans le schéma *ci-dessous*)[1117], à optimiser les trajets et même la consommation de fuel des porte-containers[1118].

[1115] Voir AI & Machine Learning are taking over the shipping industry par Ultimate Maritime Logistics, mai 2019.

[1116] ClearMetal (2014, USA, $31M) optimise avec sa plateforme CDX la gestion du transport maritime pour les clients industriels et fait appel à de nombreuses briques de machine learning.

[1117] Source du schéma : A Multi-task Deep Learning Architecture for Maritime Surveillance using AIS Data Streams par Duong Nguyen et al, IMT Atlantique Brest, 2018 (10 pages).

[1118] Avec l'application de SailRouter (Pays-Bas) qui tient compte de la météo, des caractéristiques des porte-containers et de leur charge. Voir aussi Utilization of a deep learning-based fuel consumption model in choosing a liner shipping route for container ships in Asia par LinhBui-Du et al, 2020 (11 pages) ainsi que Prediction of vessel propulsion power using machine learning on AIS data, ship performance measurements and weather data par Q Liang, 2019 (13 pages). C'est aussi la spécialité de **We4Sea** (2016, Pays-Bas, 400K€) qui réalise du suivi de performance du système de propulsion des navires.

Le machine learning est aussi utilisé pour optimiser l'activité des ports d'accueil des porte-containers[1119], pour prévoir le coût des assurances couvrant les conteneurs endommagés[1120] ou pour optimiser la gestion d'une flotte d'un armateur[1121].

Dans cet ordre d'idée, **Odyn** (2015, France) utilise le machine learning pour comprendre la dynamique des océans et aider à piloter la navigation. Il est intégré dans la solution logicielle sur tablette SeaWaze qui est proposée aux sociétés maritimes et aux affréteurs maritimes.

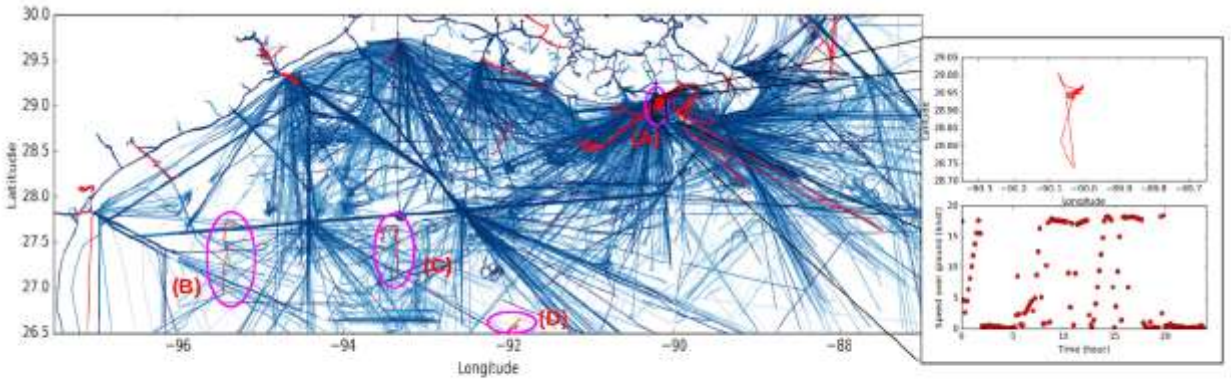


Fig. 4. Detection of abnormal behaviors using global thresholding (Gulf of Mexico dataset). Blue: tracks in the training set (which itself may contain abnormal tracks); red: abnormal tracks detected in the test set. We highlight four examples: a track with an abnormal speed pattern **(A)** ii), two tracks with abnormal trajectory shapes from others' in the same region **(B,C)** iii) a track in a low-density area (abnormal zone) **(D)**.

Enfin, l'IA fait aussi son arrivée dans la robotisation de certains navires. Dans cette catégorie, il faut distinguer les navires qui seraient entièrement autonomes, sans équipage, et ceux qui seraient opérés par des humains à partir de postes de commande à terre comme le sont les drones militaires lanceurs de missiles. Les deux catégories seront probablement hybridées en fonction des besoins[1122]. Le pilotage à distance est l'objet du projet ROSS (Remotely Operated Service at Sea) de SeaOwl lancé en France en 2018[1123].

Quel est l'intérêt des navires autonomes ? Ils permettraient en théorie de réduire les erreurs d'origine humaine qui seraient la cause de 80% des accidents maritimes. Ils serviraient surtout à réduire les coûts opérationnels liés à l'équipage, qui peuvent aller jusqu'à 30% d budget opérationnel. L'économie d'équipage permet aussi de construire des navires sans lieux de vie, donc moins chers en structure ainsi qu'en coût énergétique pour la propulsion.

De nombreux pays « marins » investissent dans les navires autonomes. C'est notamment le cas du Japon[1124]. Les expérimentations ont démarré dès 2012 en Europe avec des navires prototypes, avec notamment un ferry entièrement autonome, le Falco du Finlandais **FinnFerry**, armé en 1993, et équipé avec l'aide de **Rolls Royce** et testé en 2018 (*ci-dessous* à gauche).

Un porte-container côtier autonome est en construction, le norvégien **Yara Birkeland**, avec une capacité de 120 TEU, équivalent à 60 containers de 32 pieds (*ci-dessous* à droite). Propulsé électriquement via des batteries alimentées en partie par énergie solaire, il doit commencer sa navigation autonome à partir de 2022.

[1119] Le Singapourien PSA Maritime fait ainsi appel à EY pour optimiser la logistique du port de Singapour. Ce genre de solution est aussi déployé en Allemagne. Voir German port first to adopt ML in shipping container management par Rachel England, juillet 2020.

[1120] Voir la thèse Optimizing Shipping Container Damage Prediction and Maritime Vessel Service Time in Commercial Maritime Ports through High Level Information Fusion de Ashwin Panchapakesan, 2019 (170 pages).

[1121] C'est ce que propose Nautilus Labs (2016, USA, $13M).

[1122] Voir The Incredible Autonomous Ships Of The Future: Run By Artificial Intelligence Rather Than A Crew par Bernard Marr, Forbes, 2019.

[1123] Voir France demonstrates its autonomous shipping prowess par Sam Chambers, septembre, 2020

[1124] Voir Japan says it's ready to launch AI-driven autonomous ships by 2025, septembre 2020.

Mais comme pour tous les systèmes autonomes, la mise au point semble plus difficile que prévu[1125]. Les défis technologiques sont voisins de ceux des voitures autonomes : qualité des informations collectées par les capteurs, leur interprétation et les prises de décision associées.

Comme pour les voitures autonomes, la réglementation s'adapte, à commencer par l'établissement de règles pour l'évaluation de ces navires, comme ce que propose l'Union Européenne[1126]. Il existe aussi une classification du niveau de l'autonomie des navires qui créé une gradation allant jusqu'à l'autonomie complète, comme pour les voitures autonomes de niveau 5[1127].

Figure 2: Autonomy levels proposed by Lloyd's Register of Shipping (UK)

Autonomy Level (AL)		Details
AL0	Manual - No autonomous function	—
AL1	On-ship decision support	All actions at the ship level are taken by a human operator, but a decision support tool can present options or otherwise influence the actions chosen, for example DP Capability plots and route planning.
AL2	On and offship decision support	All actions at the ship level taken by human operator on board the vessel, but decision support tool can present options or otherwise influence the actions chosen. Data may be provided by systems on or off the ship, for example DP capability plots, OBM configuration recommendations, weather routing.
AL3	"Active" human in the loop	Decisions and actions at the ship level are performed autonomously with human supervision. High impact decisions are implemented in a way to give human operators the opportunity to intercede and override them. Data may be provided by systems on or off the ship.
AL4	Human on the loop – operator's primacy	Decisions and actions are performed autonomously with human supervision. High impact decisions are implemented in a way to give human operators the opportunity to intercede and override them.
AL5	Fully autonomous (& rarely supervised)	Unsupervised or rarely supervised operation where decisions are made and actioned by the system, i.e. impact is at the total ship level.
AL6	Fully autonomous (& with no supervision)	Unsupervised operation where decisions are made and actioned by the system, i.e. impact is at the total ship level.

Source: Compiled by MGRR based on a report by Lloyd's Register of Shipping (UK), "Autonomous Shipping 2019 and Beyond" (January 2019), and a Japanese-language report by Japan's Ministry of Land, Infrastructure, Transport and Tourism, "Challenges and Direction of Initiatives for Autonomous Ships" (December 2017).

C'est du côté de l'océanographie qu'il faut aussi regarder, avec le **Mayflower** ou MAS (Mayflower Autonomous Ship), construit par l'IRT de recherche marine ProMare et équipé d'un serveur IBM Power AC922 comprenant des GPGPU Nvidia V100[1128]. Le pilotage autonome est développé par la startup **Marine AI** (UK). Le Mayflower doit tenter de traverser l'Atlantique au printemps 2021 de manière entièrement autonome. C'est un catamaran alimenté par énergie solaire.

[1125] Voir Yara Birkeland tipped to launch late next year par Jason Jiang, novembre 2020.

[1126] Voir Maritime Autonomous Ships and Shipping - EU Operational Guidelines on trials of Maritime Autonomous Surface Ships (MASS), Union Européenne, 2020 ainsi que Autonomous shipping and its impact on regulations, technologies, and industries par Mingyu Kim, juin 2020 (10 pages).

[1127] Schémas trouvés dans Maritime autonomous surface ships : development trends and prospects par Koji Wariishi, 2020 (8 pages) et dans Artificial Intelligence and the Era of Autonomous Shipping par Baibhav Mishra, janvier 2020.

[1128] Voir Les atouts des navires autonomes par IBM, 2020.

Marine militaire

Du côté militaire, la marine américaine a lancé un programme de navires autonomes sous l'égide de la DARPA. Le **Sea Hunter** est pilote a été lancé en 2016 sous la forme d'un trimaran de 40 m et 135 tonnes (*ci-dessous* à gauche). Un programme de production de série a été lancé fin 2020 pour un montant de $300M. Le Sea Hunter est un « Medium Displacement Unmanned Surface Vehicle » (MDUSV)[1129]. L'autonomie de ce genre de navire est contrainte par la propulsion diesel. Il est donc dans un premier temps réservé à de la surveillance côtière. Un premier prototype doit être livré en 2023 avant le lancement d'une production en série de ces navires qui serviront à la reconnaissance et au renseignement. Cela permettra d'éviter des saisies dommageables comme celle du fameux navire de surveillance Pueblo qui était géré par la NSA, en 1968 dans les eaux nord-coréennes.

L'US Navy annonçait aussi lancer en 2020 un programme d'une dizaine de « Large Unmanned Surface Vessels » (LUSV) d'un tonnage plus important, 2000 tonnes, et construit par Lockheed Martin, pour un budget de $2B. Elle prévoit aussi de lancer un sous-marin autonome, de type « Extra-Large Unmanned Undersea Vessel » (XLUUV).

Les Chinois ne sont pas en reste. Ils lançaient en 2020 un prototype de petit patrouilleur devant devenir entièrement autonome, le JARI USV faisant 20 tonnes et 15 mètres de long. Il est équipé pour des missions défensives tout azimut (contre sous-marins, navires de surface et anti-aérien) via huit lance-missiles intégrés dans la coque, un lance torpilles et un canon. Le navire a une autonomie de 500 miles nautiques (*ci-dessus* à droite).

Bâtiments et Travaux Publics

Le marché du BTP comprend de nombreuses sous-catégories avec celui du gros-œuvre et du second-œuvre dans la construction de bâtiments d'habitation ou dans le tertiaire, celui de la promotion immobilière puis celui du génie civil pour les gros ouvrages d'art ainsi que pour la construction de routes.

[1129] Voir U.S. Navy Moves Forward With its Autonomous Vessel Program, décembre 2020.

En France, cette activité représente plus de 170 Md€ de CA et plus de 1,4 millions d'emplois répartis dans environ 530 000 sociétés allant de l'artisan aux grands acteurs tels que Vinci et ses filiales Eurovia et BATEG, le groupe Bouygues (Colas, Bouygues Construction, Bouygues Immobilier), Spie Batignolles, Eiffage, Nexity et Quartus.

Les grandes tendances du secteur comprennent la construction écologique, la réalité augmentée, les robots constructeurs, les drones, les logiciels de suivi de chantier, les méga-projets et le BIM (Building Information Management)[1130]. Quatre de ces tendances ont un lien direct avec l'IA.

La **Fédération Française du Bâtiment** s'est emparée de ce sujet en produisant un rapport en 2019 qui inventorie convenablement la large panoplie des usages actuels et potentiels de l'IA dans ses métiers[1131] !

Ces usages couvrent tous ces métiers mais avec une grande variation selon la taille des structures. J'ai observé une loi empirique selon laquelle les innovations perfusent moins rapidement dans les marchés b2b très fragmentés.

Les premiers industriels du secteur à adopter les nouvelles technologies sont plutôt les plus gros d'entre eux. Les artisans et TPE/PME n'ont pas les mêmes capacités d'adaptation, sauf lorsqu'il s'agit d'utiliser des outils numériques génériques (télécoms, ordinateurs, smartphones).

Construction

Le secteur de la construction n'a pas attendu la mode du deep learning de ces 10 dernières années pour se mettre à l'IA. Il était déjà un assez gros consommateur de systèmes experts depuis les années 1980, mais avec plus ou moins de bonheur.

En effet, nombre d'expériences de cette époque n'ont pas forcément porté leurs fruits. Comme dans de nombreux domaines, je suis donc allé à la pêche aux informations en suivant plusieurs fils d'Ariane, aidés par mon ami le moteur de recherche.

Des chercheurs de l'université Zhejiang de Hangzhou évoquent de nombreux cas de modélisations complexes faisant appel à des systèmes experts et de la logique floue[1132]. La logique floue est utilisée pour la modélisation d'expertise informelle. On y dénombre des applications de prévision de la solidité des routes, de l'optimisation de trajets, la prévision du comportement d'ouvrages de génie civil et de la maintenance préventive. Des outils de planification intègrent les risques et incertitudes dans les chantiers. Les solutions font appel à des algorithmes évolutionnaires, génétiques et des réseaux multi-agents. Tout cela, avant l'explosion cambrienne du deep learning ! La planification de constructions pouvait aussi exploiter des systèmes experts[1133].

Le deep learning a ensuite fait son apparition dans des applications très spécifiques comme pour prévoir la solidité de structures en béton en utilisant un réseau de neurones **FPNN** pour Fuzzy Polynomial Neural Networks[1134]. Des modèles de prévision de résistance du béton au feu par réseau de neurones ont aussi été créés[1135].

[1130] Voir Les 7 tendances qui vont révolutionner le secteur du BTP en 2019, mai 2019.

[1131] Voir Intelligence artificielle et bâtiment, Fédération Française du Bâtiment, mai 2019 (40 pages).

[1132] Dans Artificial Intelligence in Civil Engineering, 2012 (22 pages).

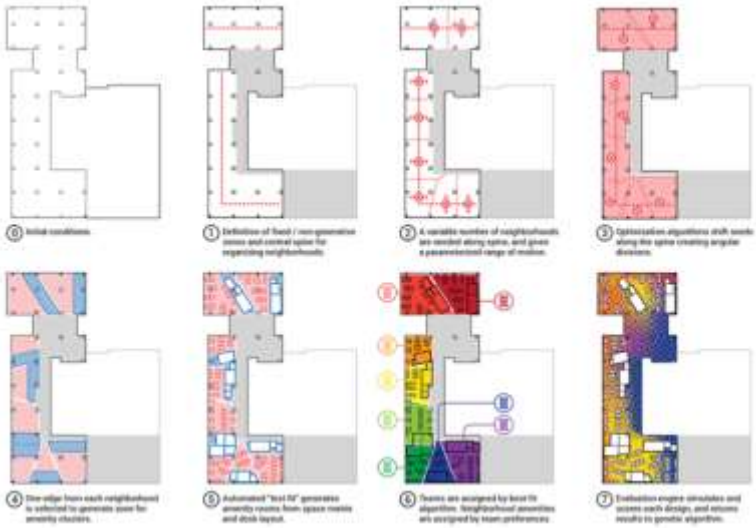
[1133] Voir Application of Artificial Intelligence in Construction Management (14 pages).

[1134] Voir Rule-Based Fuzzy Polynomial Neural Networks in Modeling Software Process Data, 2003 (11 pages).

[1135] Voir Application of Artificial Neural Networks in Civil Engineering, 2014 (7 pages).

Dans un inventaire récent d'usages de l'IA dans le BTP[1136], j'ai trouvé **Neuro-modex**, un modèle permettant d'optimiser les décisions de construction en fonction du lieu, de l'environnement, des conditions d'emplois, des caractéristiques du projet et de ses risques, de outils de planification avancée et des outils d'optimisation financière de la réfection de la chaussée.

Encore plus récemment, est apparue une solution de détection de failles structurelles dans les ouvrages de génie civil. Elle n'a visiblement pas été employée sur l'autoroute de Gênes en Italie ! C'est à base de logique floue et de réseaux de neurones simples[1137].



Des équipes d'**Autodesk** présentent des outils de design génératifs à base de GANs permettant de définir l'organisation des bureaux. C'est le projet MaRS Generative Design[1138]. Ca fait rêver puisque cela sert à aménager des open spaces !

SmartVid.io (2015, USA, $21,8M) propose une solution de vidéosurveillance de chantiers qui permet de s'assurer que les ouvriers travaillent dans de bonnes conditions de sécurité. Elle va par exemple détecter les ouvriers qui ne sont pas dotés des bons équipements de sécurité (casque, gants) et détecter d'autres facteurs de risque (eau, positions instables en hauteur, …) jusqu'à faire des prédictions sur des incidents probables. Les briques logicielles à assembler pour faire fonctionner tout cela correctement ne sont pas pour autant présentées sur le site de la startup[1139].

Buildots (2018, Israël, $16M) propose une autre solution d'inspection de chantiers qui exploite une caméra 360° placée au-dessus de casques d'ouvriers ou chefs de chantiers qui permettent d'analyser l'environnement à raison de deux images par secondes. Les images captées, une fois interprétées, permettent de générer des états de suivi du chantier et d'analyser éventuellement des malfaçons a posteriori (vidéo).

Le secteur de la construction est friand de robots et autres exosquelettes pour à la fois accélérer les chantiers et réduire la pénibilité du travail tout comme celle des accidents de personnes[1140]. Plusieurs startups et industriels se sont donc lancés dans la création de robots divers allant de l'automate téléguidé au robot véritablement autonome, bien plus rare[1141].

On assiste à un phénomène voisin de celui que nous évoquerons plus loin dans l'agriculture, à savoir, l'émergence de nombreux prototypes qui ne sont souvent pas industrialisés. Soit ils ne sont pas au point, soit leur équation économique n'est pas encore satisfaisante, soit les deux à la fois.

[1136] Voir Artificial Intelligence as a Tool in Civil Engineering – A Review, 2017 (4 pages).

[1137] Voir Comparison of Artificial Neural Networks and Fuzzy Logic Approachs for Crack Detection in a Beam Like Structure, 2018 (17 pages).

[1138] Voir The Promise of Generative Design, 2017, et détaillé dans The Rise of the AI: Impact of AI and Machine Learning in Construction (12 pages).

[1139] Voir 10 Ways AI Is Improving Construction Site Security par Louis Columbus, 2020 qui évoque les différentes solutions à base d'IA de sécurisation de chantiers.

[1140] Voir Robotics in Construction, 2016 (87 pages).

[1141] Voir 8 Startups Building Robotic Construction Workers, novembre 2017.

On peut citer le cas de **Doxel** (USA, $4,5M) qui utilise un petit robot chenille et des drones équipés d'un LiDAR et de caméras pour inspecter les chantiers de construction autant en intérieur qu'en extérieur (vidéo). L'outil logiciel évalue la progression des travaux et est censé détecter les malfaçons. La startup affiche pour l'instant trois clients : Kaiser Permanente (San Diego), Lucas Museum (Musée de Georges Lucas en Californie) et Sutter Health (hôpitaux, également en Californie).

Les barrières à l'adoption des robots dans ce secteur d'activité sont leurs coûts d'acquisition, d'installation, d'opérations et de maintenance, la fragmentation du marché et des tâches à robotiser qui génèrent un besoin d'une grande variété de robots pour automatiser des chantiers[1142], les difficultés de leur mise au point, l'incompatibilité avec les pratiques usuelles dans les chantiers et en dernier ressort, l'acceptation des nouveaux outils par les ouvriers[1143].

Avant la construction, il a la démolition ! On trouve des robots de démolition qui sont en fait des machines téléguidés chez **Brokk** avec son modèle 180 (*ci-dessous* à gauche) et **Husqvarna** et son DXR (*ci-dessous*, au milieu) et, plus intrigant, le **ERO** conçu en 2013 par un étudiant en Suède et prototypé par le fabricant de compresseurs **Atlas Copco** *(ci-dessous, à droite)*. Ce robot utilise un jet d'eau sous pression pour détruire les structures de murs en béton et récupérer les débris pour les recycler. Ces débrits sont aspirés et servent ensuite de matériau de construction comme gravier dans la composition de bétons spéciaux. Il ne semble cependant pas que ce robot ait été produit en série. On n'en trouve pas de vidéo le montrant en état opérationnel. C'est un cas classique de communication difficile à interpréter ! Les images présentées dans la communication d'ERO et d'Altas Copco ne sont que des modèles de synthèse en 3D ! Dommage.

Nous avons déjà évoqué dans la rubrique des exosquelettes les cas d'**Ekso Bionics** et du français **DB3D** qui sont aussi présents dans le marché de la construction. Ce dernier fournit des produits expérimentés par Colas en France pour la construction de routes et la réfection des chaussées[1144].

[1142] Robots de démolition, de terrassement, de levage, de soudure, de pose de ciment, de parage de surfaces, de peinture, de construction de routes, etc.

[1143] Voir An investigation into the barriers to the implementation of automation and robotics technologies in the construction industry, de Rohana Mahbub, 2008 (303 pages) qui étudie les obstacles à la robotisation de la construction en Australie, au Japon et en Malaisie. L'étude a 10 ans mais ses conclusions semblent être toujours d'actualité.

[1144] Voir Colas lance le déploiement des exosquelettes sur ses chantiers, mars 2018.

Built Robotics (2016, USA, $48M) développe des robots tracteurs dédiés à l'excavation de terrains et qui exploitent les mêmes capteurs que les voitures autonomes. Le fondateur est un ancien ingénieur de Google qui n'y a pas travaillé plus d'un an en 2010. Le principe de fonctionnement est voisin de celui d'une tondeuse de gazon autonome : l'opérateur délimite les contours du terrain à déblayer et la zone où déposer la terre, et le robot se débrouille tout seul ensuite (vidéo).

Fastbrick Robotics (2015, Australie, $36,6M) a créé le robot Hadrian X qui dépose des briques ou parpaings pour construire des bâtiments. Il peut ainsi construire un bâtiment de un à deux niveaux en quelques jours (vidéo). Caterpillar a investi $2M dans cette société et l'Arabie Saoudite a signé un MOU pour construire 50 000 maisons avec l'enfin d'ici 2022. Petit détail observable : le robot ne dépose pas de mortier sur les briques pendant leur pose. Ca fait un peu désordre ! Avec l'attribution de la citoyenneté à Sophia par le pays, cela fait la paire !

Construction Robotics (2008, USA, $75K) développe aussi un robot de pose de briques qui lui, sait poser le mortier autour. Le SAM100 (Semi-Automated Mason) fait gagner 80% du temps de travail du maçon et ne le remplace donc pas entièrement. C'est un « cobot » qui peut déposer 2000 briques par jour contre 400 pour un maçon. Petit détail à noter : il ne fonctionne que pour des murs droits. Il n'aime pas encore les angles !

Ce qui est gênant pour de la construction d'habitations et le rendra plus utile pour la construction de bâtiments industriels de grande taille (vidéo).

Cazza (2016, USA, $2M) a créé le X1, un robot-imprimante 3D comme il en existe déjà quelques-uns. La machine est mobile grâce à ses chenilles. Il imprime des structures verticales ou coule des dalles horizontales. Où est l'IA dans un tel robot ? Il doit normalement disposer de capteurs et de caméras pour s'orienter. Au minimum doit-il donc disposer d'outils de reconnaissance d'images pour se déplacer (vidéo).

Endless Robotics (2015, Inde, $100K) a créé le robot WALT qui peut peindre les murs 30 fois plus rapidement qu'un peintre doué, mais a tout de même besoin d'être supervisé par un humain. Et comme nombre de robots cités ici, c'est encore un prototype (vidéo). Et puis, dans la construction, il y a aussi les robots expérimentaux de **Schaft** (2012, Japon)[1145], abandonnées par Google en 2018 après en avoir fait l'acquisition en 2014, et le HRP-5P Humanoid Robot du laboratoire japonais AIST (vidéo). Il est capable de poser des planches. Il mesure 1,82 m et pèse 100 kg[1146].

OpenSpace (2017, USA, $33,5M) propose une solution à base de reconnaissance d'image qui permet de recréer facilement un modèle 3D d'un site de construction grâce à des captures video 360° obtenues par des cameras attachées aux casques de chantiers, couplées à de l'annotation des points clés. Le tout est associé à une fonction de BIM (Building Information Management) permettant de comparer l'état du chantier avec les plans 3D (vidéo). Cela facilite le reporting d'avancement du chantier. **Disperse.io** (2015, UK, 1,3M£) est une autre startup qui propose le même genre de solution.

Arcure (2009, France, 1,5M€) commercialise Blaxtair, qui associe des capteurs de vision 3D stéréoscopique et de l'IA embarquée. Elle permet la détection périmétrique de piétons pour sécuriser les déplacements des véhicules industriels. La solution a été développée en collaboration avec le CEA. Leur déclinaison dans le capteur Oméga est adaptée à la vision pour les usines connectées et la robotique industrielle d'extérieur ou d'intérieur.

Seismic.ai (2016, Israël) utilise les milliers de séismes indétectables pour entraîner son IA pour prévoir plus en avance de gros séismes.

Pour terminer sur l'usage de l'IA dans les chantiers, on pourrait intégrer la dimension humaine. Le recrutement d'ouvriers de chantiers peut dans une certaine mesure exploiter les outils évoqués dans la partie RH. La vidéo-surveillance permet de vérifier le respect des règles de sécurité. Certaines solutions permettent par exemple de s'assurer du respect de l'accès aux zones sécurisées. Les drones d'inspection peuvent aussi surveiller les chantiers et leur avancement.

Immobilier

Le marché de l'immobilier est aussi potentiellement friand d'IA, qu'il s'agisse d'améliorer les étapes de la création, de la planification des offres et de leur commercialisation. C'est un secteur d'activité où les grands promoteurs s'activent mais où sévissent pas mal de startups, en particulier pour rapprocher l'offre de la demande de logements aussi bien à l'acquisition qu'à la location[1147].

En voici quelques-unes.

[1145] Le robot de Schaft avait gagné le DARPA Robotic Challenge en 2014 (vidéo). Dans une autre vidéo datant également de 2014, on le voit réaliser diverses tâches de chantier en pièces détachées (vidéo).

[1146] Voir Humanoid robot prototype HRP-5P capable of the same hard work as human being developed, septembre 2018. Le robot est encore « mono-tâche » pour l'instant.

[1147] Voir Les enjeux et défis de l'intelligence artificielle dans la filière de l'immobilier, Xerfi, 2019

Skyline (Israël, $21M) fait du benchmarking de biens immobiliers multicritères pour les investisseurs, pour leur permettre de dénicher les bonnes affaires. Les sources d'informations utilisées sont nombreuses (130) avec plus de 10 000 attributs par actif.

Enodo (2016, USA, $2,5M) propose un outil équivalent au précédent pour évaluer les investissements immobiliers prometteurs en termes de rendement.

GeoPhy (2014, Pays-Bas, $33M) utilise le machine learning pour évaluer la valeur de biens immobiliers basée sur différentes sources d'information y compris de l'imagerie satellite[1148].

REX Real Estate (2014, USA, $25,5M) utilise une forme d'apprentissage par renforcement pour améliorer l'efficacité du ciblage publicitaire en ligne d'annonces immobilières. Il analyse les paramètres communs de segmentation des premiers utilisateurs qui cliquent sur l'annonce pour renforcer la visibilité de l'annonce sur le segment qu'ils représentent.

First.io (2014, USA, $7,35M) propose une solution logicielle de prévision des déménagements des foyers, pour leur proposer des offres au bon moment. La société suit 700 paramètres sur 214 millions de personnes aux USA, à la fois des paramètres personnels comme ceux qui sont liés aux modifications de l'environnement (modifications de l'aménagement des quartiers, ouvertures/fermetures d'écoles, etc). C'est finalement une sorte de Cambridge Analytica adapté à l'immobilier !

OJO Labs (2015, USA, $26,5M) a créé un chatbot de relation client pour les agences immobilières. Il permet aux utilisateurs de trouver des biens qui correspondent à leurs critères de choix. Au passage, leur système reconnaît les caractéristiques des biens via les photos qui l'accompagnent.

Appartenir (France) est un service recherche immobilière qui envoie des alertes sur des annonces qui répondent à ses critères de recherche. L'IA utilisée par la startup n'est pas précisée car on n'en a pas forcément besoin pour faire des recherches multi-critères dans une base d'annonces, qui plus est lorsque sa taille est relativement réduite. Les annonces sont récoltées sur divers sites qui les publient.

CompareAgences (2012, France) facilite la relation entre agents immobiliers et particulier dans le cadre de la vente de biens. La startup emploie 12 personnes et génère 200 000 visiteurs uniques par mois. 1000 agences immobilières sont intégrées en France. Le tout est à base de machine learning, sans plus de précisions.

Smart city

Le concept de Smart City a été créé il y a une dizaine d'années et promu notamment par IBM dans le cadre de son initiative Smart Planet (en 2008). De nombreux fournisseurs de technologie ont embrayé le pas, comme Cisco puis Nvidia. Le marché du secteur public et des villes étant important pour eux, le concept de Smart City était et reste un moyen d'adopter un discours commercial global pour approcher les collectivités locales.

Diverses villes ont lancé des initiatives de ce genre, telles que Barcelone (avec Cisco), Nice (encore avec Cisco) et Amsterdam, généralement axées sur les transports. Un tas de grandes villes du monde ont emboîté le pas, clamant haut et fort leurs plans de Smart City. Le *Smart City washing* est devenu un outil politique puissant.

Les diverses solutions de Smart City portent sur l'organisation des moyens de transports, de l'éclairage, de la fourniture d'énergie[1149] et eau, de la sécurité, voire des systèmes de santé. Elles portent aussi sur la maintenance préventive des infrastructures.

[1148] Voir Using AI for Commercial Real Estate Valuations, janvier 2019.

A chaque besoin sa solution et ses sources de données. Le concept de Smart City reste cependant flou car son approche systémique reste assez théorique et difficile à mettre en œuvre[1150].

En effet, sauf à être lancé dans une dictature, la plupart des grandes villes connaissent un développement généralement non coordonné et l'existant est difficile à moderniser d'un coup de baguette magique. Ce d'autant plus que la Smart City est souvent une grande dévoreuse de données issues des infrastructures et des activités des citadins avec des conséquences sur la vie privée qui sont encore mal évaluées.

Des projets symboliques peuvent être lancés, tels que le **Toronto Waterfront** conduit par **Sidewalk Labs**, une filiale d'Alphabet créée en 2015[1151]. Il devait s'appuyer sur l'usage de véhicules autonomes et de l'habitat intelligent, mais sur une zone de surface limitée à quelques hectares, en bordure du lac Ontario devant être entièrement reconstruite (*ci-dessous*).

Le projet est allé jusqu'à l'étape de l'étude de faisabilité et de la consultation publique des besoins et idées[1152] (vidéo). Il a généré son lot d'opposition des citoyens actifs dans la ville, notamment autour des questions de préservation de la vie privée[1153]. Ce projet devait permettre de créer un « living lab » des idées constitutives de la Smart City et d'en conduire des évaluations objectives, en particulier pour ce qui est de l'usage de véhicules autonomes qui sera probablement plus sûr dans un environnement où ils cohabitent peu avec des véhicules conduits traditionnellement. En mai 2020, le projet était abandonné, la raison officielle étant les difficultés économiques résultant de la pandémie covid-19.

En pratique, la « ville intelligente » est rarement une ville intégrée. C'est une ville qui met en œuvre des services utilisant des technologies disparates pour améliorer la vie de ses habitants, réduire le coût de ses infrastructures et son empreinte énergétique.

[1149] Voir par exemple le projet français AMOEBA de l'IRIT (Institut de Recherche en Informatique de Toulouse), qui vise à trouver les corrélations entre prévisions de consommation d'énergie et la production d'énergie, décrit dans Use Cases of Pervasive Artificial Intelligence for Smart Cities Challenges (52 slides) et Use Cases of Pervasive Artificial Intelligence for Smart Cities Challenges (7 pages).

[1150] Voir par exemple City Brain, a New Architecture of Smart City Based on the Internet Brain 2017 (8 pages), un document chinois qui cherche à créer une analogie conceptuelle entre le fonctionnement d'une ville intelligente et le cerveau.

[1151] A smarter smart city - An ambitious project by Alphabet subsidiary Sidewalk Labs could reshape how we live, work, and play in urban neighborhoods de Elizabeth Woyke , février 2018.

[1152] Voir la présentation Roundtable 3 qui décrit assez bien l'étendue du projet, août 2018 (118 slides).

[1153] Voir Le projet SmartCity de Google à Toronto repart par Frédéric Charles, novembre 2019.

D'un point de vue pratique, la Smart City s'appuie sur des briques technologiques communes que sont les capteurs et les objets connectés, les infrastructures de télécommunications pouvant comprendre des réseaux M2M (Sigfox, LoRA) et des applications construites autour de « big data » et de machine learning pour exploiter les données générées par les capteurs. Le machine learning intervient naturellement pour faire des analyses de données, des segmentations, des détections d'anomalies et des prévisions.

Les applications à base d'IA qui reviennent le plus souvent dans les villes intelligentes concernent l'**optimisation des transports**, souvent routiers, principalement pour les décongestionner et en assurer la sécurité. L'étape ultime de la transformation sera le passage aux véhicules autonomes mais ce n'est pas encore à l'ordre du jour dans la majorité des projets.

La vidéo-surveillance est de plus en plus utilisée pour l'optimisation du trafic routier[1154] et bénéficie des nombreuses avancées dans la reconnaissance d'image via le deep learning. L'exploitation des images de vidéo-surveillance permet d'évaluer le trafic, la vitesse des véhicules, les embouteillages, la détection d'accidents, et dans certains cas, le pistage dans la durée de véhicules suivis par la police. On peut aussi exploiter les GPS des mobiles pour prévoir les mouvements de foule[1155].

Nvidia organise même depuis 2017 un « AI City Challenge » qui est focalisé sur la vidéo-surveillance, ce qui s'explique par leur offre qui associe des chipsets pour serveurs (Volta V100) et pour équiper les caméras, les Jetson TX1. Des dizaines de projets sont sélectionnés et financés qui font avancer graduellement les techniques d'analyse d'images vidéo.

RapidFlow Technologies (2015, USA) développe des systèmes de signalisation (feux) exploitant des caméras pour ajuster leur programmation dynamiquement en fonction du trafic et de manière coordonnée entre feux. Le système est déployé à Pittburgh, réduisant les temps d'attente de 40% d'après la startup.

Vivacity Labs (2015, UK, £3,3M) permet d'optimiser en temps réel le trafic dans la rue et le stationnement par analyse d'images. Il peut être couplé aux systèmes de régulation de trafic (vidéo).

RoadBotics (2016, USA, $11,4M) utilise des smartphones pour capter visuellement l'état des routes et indiquer les travaux de réparation à lancer (nids de poule, chutes d'arbres, obstacles).

Xaqt (2015, USA) exploite les données issues des capteurs intégrés dans les routes, les caméras et les données météo pour prédire l'apparition de nids de poule dans la chaussée avec une précision de 85% ! Cela permet de déclencher les réparations plus rapidement. Ou pas, selon les habitudes des services de la voirie !

Near Space Labs / Swiftera (2017, USA, $1,5M) exploite de l'imagerie issue d'un ballon d'observation créé sur mesure pour cartographier les villes et leur permettre de gérer leurs programmes d'urbanisation (vidéo).

Qucit (2014, France, 1,7M€) propose des solutions aux collectivités locales pour optimiser la gestion de vélos en libre-service, améliorer l'efficacité du contrôle du stationnement en optimisant le parcours des agents de verbalisation, cartographier des points d'attention et de satisfaction d'un territoire et enfin, pour l'optimisation du trafic routier et des autoroutes (vidéo). Autant de solutions ad-hoc qui exploitent de la donnée et du machine learning.

Datategy (2016, France) est une startup qui lutte contre la fraude dans les transports et les parkings. Elle récupère toutes les données imaginables pour identifier les zones où la fraude est la plus forte et y envoyer les contrôleurs avec leurs terminaux de verbalisation. Cela utilise force machine learning. Dans la même veine **Vimoc** (2012, USA, $2,4M) optimise la gestion des parkings.

[1154] Vu dans Video Analytics for AI City Smart Transportation de Ming-Ching Chang, 2017 (41 slides).

[1155] C'est l'objet d'un projet de Microsoft Research publié en janvier 2017 : Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction (7 pages).

L'IA est aussi employée sur les autoroutes chez Vinci comme chez son concurrent SANEF.

Chez **Vinci**, le système Cyclope analyse les flux vidéo provenant de caméras existantes et réalise de la classification automatique des véhicules pour valider les tarifs de péage en cas d'ambiguïté et détecte les anomalies de comportement sur autoroute comme les véhicules arrêtés, les piétons ou les mauvais sens de circulation[1156]. Le dispositif a déjà été expérimenté sur les autoroutes du Sud de la France (Escota).

Le réseau **SANEF** (autoroutes du Nord et de l'Est de la France) s'apprête à moderniser ses péages avec l'analyse des images de caméras et de LiDAR pour classifier les véhicules, tout en lisant les plaques d'immatriculations. C'est couplé à de l'analyse de trafic servant à repérer des anomalies et à remonter des alertes. Le système entrait en production début 2019 dans deux grandes gares de péage de l'autoroute A4[1157].

Utilities

Les *utilities* sont les fournisseurs d'électricité, de gaz et d'eau ainsi que les gestionnaires des déchets. Ils sont de grands utilisateurs potentiels de collecte de données et d'exploitation de l'IA pour optimiser leurs ressources de production et leurs infrastructures. Leur objectif est d'adapter dynamiquement la production à la consommation ou de modérer cette dernière si besoin est.

Il s'agit aussi de réduire leur empreinte énergétique et de réaliser de la maintenance prédictive de leurs infrastructures de production et de transport.

Enfin, accessoirement, il s'agit aussi d'améliorer le service rendu aux clients. En général, le bon service se manifeste quand tout fonctionne sans que l'on ait à s'en soucier.

Les utilities exploitent l'IA dans différents autres domaines. Chez **EDF**, c'est la comptabilité des 24 000 fournisseurs et de la communication par emails avec eux qui est accélérée grâce à un assistant qui prépare les réponses à environ 60% des demandes. Chez **Engie** Home Services, qui gère l'entretien et le dépannage des systèmes de chauffage et de climatisation du grand public, la startup Vekia utilise l'IA pour améliorer sa gestion de stocks de pièces détachées[1158]. Chez **BP**, la productivité des raffineries est améliorée grâce à la combinaison de capteurs et de systèmes de simulation[1159].

Total Direct Energie a créé un chatbot permettant aux particuliers de suivre leur consommation énergétique[1160]. **IBM** promeut l'usage de ses données météo issues de l'acquisition de The Weather Channel pour conseiller les investisseurs dans les énergies renouvelables.

Maintenance prédictive

C'est le domaine où la créativité dans l'usage de l'IA semble être le plus grand et elle s'adapte aux nombreuses situations rencontrées par les utilities.

Les lignes à haute tension peuvent être inspectées par des drones qui sont pilotés automatiquement en fonction de parcours préétablis et dont les images sont aussi exploitées plus ou moins automatiquement pour détecter des défauts dans les lignes et les pylones.

Flod (2015, France) analyse leur bruit avec cymbalum-IoT pour faire de la maintenance prédictive, une solution dont les tests avaient démarré en 2016 chez Enedis.

[1156] Voir Vinci Autoroutes teste de premières applications métiers nourries à l'IA sur son réseau, février 2018.

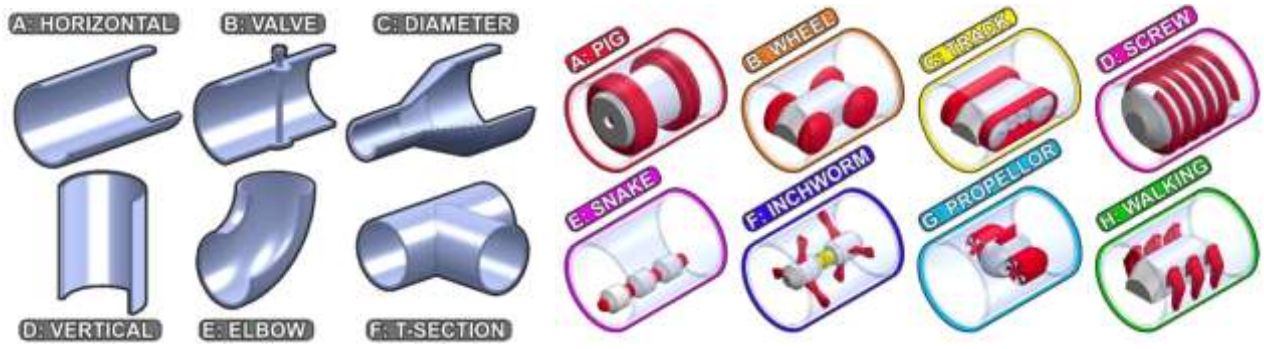
[1157] Voir SANEF : Test du 1er péage free flow avec lecture des plaques en France, de Thibaut Emme, Leblogauto, septembre 2018.

[1158] Voir ENGIE adopte l'intelligence artificielle pour sa gestion de stocks, mars 2018.

[1159] Voir How Algorithms Are Taking Over Big Oil par Christopher Helman, janvier 2019.

[1160] Retour d'expérience: Comment Direct Energie a développé son Chatbot? (4 mn). Il n'est pas certain que l'on ait un besoin fréquent de ce genre d'outil.

General Electric (USA) est l'un des nombreux industriels qui proposent des engins et services d'inspection des pipelines, comme l'échographe UltraScan Duo. Le contrôle non destructif est souvent réalisé avec des sondes dénommées « smart pigs » pour "Pipeline Inspection Gauges") qui parcourent les pipelines en les inspectant de l'intérieur. Mais on peut faire appel à d'autres engins pour les parties non linéaires des pipelines[1161].



Les techniques d'inspection sont variées avec de l'échographie, de la magnétographie ou de la mécanique de variation de géométrie et de diamètre[1162].

Elles génèrent presque toutes des images en quantité qui peuvent être analysées par des réseaux de neurones convolutifs pour identifier différents types de défauts.

Engie a développé un chatbot SMS et Skype destinée à ses techniciens de maintenance des parcs éoliens, solaires et hydrauliques. Il a été développé par la société de services Eryem sur le Bot Framework et les Cognitive Services de Microsoft puis déployé dans le cloud sur Azure.

C'est un moyen différent de naviguer dans une application de visualisation de données d'un parc d'éoliennes, solaire ou hydraulique, ou des caractéristiques d'une installation donnée (sa production, la vitesse du vent, les paramètres de fonctionnement).

Table 1. Handcrafted Features

ID	Feature	Description
1	DWATT	Raw turbine load
2	TNH	Raw turbine speed
3	MAX	Max TCs
4	MEN	Mean TCs
5	STD	Standard deviation of TCs
6	MED	Median of TCs
7	DIF	# diff b/w positive & negative TCs
8	ZR	Zero crossing
9	KR	kurtosis
10	SK	skewness
11	M3S	Max of 3-pt sum
12	M3M	Max of 3-pt median

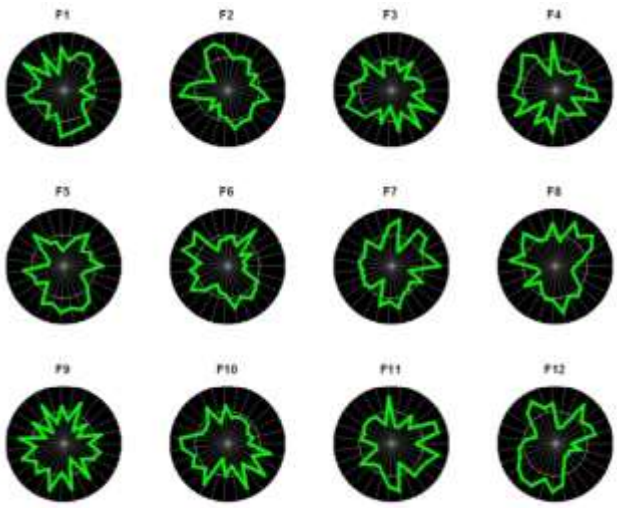


Figure 4. The 12 learned features

[1161] Voir Advances in the Inspection of Unpiggable Pipelines, 2017 (13 pages).

[1162] Neuf méthodes d'inspection en tout, inventoriées dans Critical Review of Pipeline Scale Measurement Technologies (14 pages). Voir à quoi cela ressemble dans l'édition de début 2018 dy Pipeline Technology Journal et dans Combined Crack and Metal Loss Detection Tool using Phased Array Technology, 2007 (5 pages). Evaluation of technologies to assess the condition of pipe coating on Line 5, 2018 (127 pages) évoque de son côté l'inspection des pipelines sous-marins, qui requièrent une inspection extérieure en environnement hostile sur les fonds marins.

Côté production, **General Electric** fait de la détection d'anomalies dans les turbines à gaz de centrales thermiques via la création de profils de températures en sortie de turbine. Douze paramètres sont mesurés qui génèrent une rosace de paramètres et des patterns associés détectés par un réseau de neurones[1163] (schémas *ci-dessus*).

L'IA peut aussi être utilisée dans les bassins hydrauliques, pour identifier en avance les ouvrages qui sont les plus fragiles face aux intempéries[1164].

EDF lançait sa propre startup, Metroscope, bref, une filiale, ciblant les entreprises et proposant une solution de détection d'anomalies dans les processus industriels et de maintenance prédictive. Elle a été développée en mode agile par des équipes internes d'EDF.

Voici quelques autres cas de figure, rencontrés chez diverses startups.

PlutoShift (2016, USA, $11,5M) utilise l'IA pour gérer les opérations des centres d'épuration d'eau. Le tout à base de machine learning et de beaux tableaux de bord de pilotage (*ci-contre*). C'est un éditeur de logiciel généraliste qui couvre divers besoins dans l'industrie et le secteur de l'énergie en particulier.

L'opérateur américain d'éolienne **Invenergy** exploite la solution logicielle SparkPredict issue de l'éditeur de logiciels **SparkCognition** (2013, USA, $163,6M) à base de deep learning pour déterminer les périodes de maintenance préventive des éoliennes, exploitant quatre années de données sur 100 éoliennes et 27 variables de fonctionnement. Cela rappelle qu'il faut des données de qualité pour faire du prédictif[1165]!

L'application réduit le coût moyen de chaque panne en le faisant passer de $350K à $50K et la planification des réparations intervient à plus de 2 mois d'avance (avec un coût moindre) au lieu de 2 à 3 jours avant (avec un coût plus élevé). Fin 2017, SparkCognition a été aussi sélectionné par le Département de l'Energie US pour faire de la maintenance prédictive du même genre dans les centrales au charbon[1166].

Côté éolienne, on trouve même chez **Aerones** (2015, USA, $5,2M) des drones qui en assurent le nettoyage automatique avec évidement beaucoup de traitement de l'image au programme (vidéo).

Le solaire peut aussi être de la partie. **RayCatch** (2015, Israël, $7,3M) utilise ainsi le machine learning et l'analyse des données de production d'électricité pour surveiller et optimiser les installations de panneaux solaires photovoltaïques.

[1163] Vu dans On Accurate and reliable anomaly detection for has turbine combustors : a deep learning approach, de Weizhong Yan et Lijie Yet, 2015 (8 pages).

[1164] Voir Using artificial intelligence to locate risky dams de Sarah Fecht, août 2018.

[1165] Voir aussi AI Used to Predict Failure of Wind Turbine Facility utilization rate improves from 21% to 23%, 2018.

[1166] Le fondateur de la société Amir Hussain a publié en 2018 The sentient machine, un ouvrage de vulgarisation de plus qui dépeint de futur de l'IA, allant vers l'IA générale (AGI).

Gestion du grid

Les grands utilities gèrent des réseaux complexes avec diverses ressources et de nombreux points de distribution. Les énergéticiens doivent souvent gérer l'hétérogénéité des sources de production d'énergies renouvelables (faiblement stockables) et fossiles (stockables par défaut), anticiper la production d'énergies renouvelables en fonction de la météo et des besoins de consommation pour planifier la production d'énergies d'origine fossile, piloter à grande échelle des infrastructures de production, de distribution et de consommation[1167].

Les « energy grids » peuvent aussi gérer les échanges entre réseaux de production. La notion de grid intègre aussi la supervision de la maintenance préventive, exploitant les briques d'IA vues dans la partie précédente. De grands acteurs et intégrateurs ont leur propre solution de gestion de grid, comme chez **GE** et **IBM**. Mais là encore, de nombreuses startups cherchent à rentrer dans ce marché, dont les exemples suivants.

Deepki (2014, France, 10,1M€) gère la consommation d'énergie de parcs immobiliers avec 200 000 bâtiments déjà observés pour 130 clients dont une part de la Ville de Paris et une autre d'Intermarché avec 3200 magasins et entrepôts. Au bout du compte, une économie de 5% à 10% de la facture énergétique. Les outils comprennent Collect pour consolider les données d'exploitation et Ready qui les analyse et détecte notamment les anomalies de consommation d'énergie[1168].

Metron (2013, France) détecte les sources d'économies d'énergies avec du machine learning qui est censée permettre de faire des économies d'énergie de 15%. La solution repose sur la création d'un jumeau numérique sur lequel des simulations d'économie d'énergie peuvent être réalisées[1169].

Drift (2015, USA, $7,5M) est un Enron des temps modernes qui utilise l'IA pour faire de l'échange d'énergie sur les marchés ouverts en mode peer-to-peer pour s'adapter en temps réel à l'offre et à la demande, le tout avec du trading à haute fréquence qui est justifié pour la distribution d'énergies fortement intermittentes comme l'éolien. Le service qui a démarré dans la ville de New York intègre un réseau de 3000 producteurs d'énergies renouvelables hydraulique et solaire. Le machine learning est utilisé essentiellement pour faire du prédictif, en intégrant les données climatiques et les historiques de consommation dans la zone desservie. Les clients sont les consommateurs d'énergie, aussi bien dans le logement individuel que dans les entreprises, et le prix du service est fixe.

BuildingIQ (2009, USA, $33,7M) propose une solution à base de machine learning qui permet d'optimiser et prévoir la consommation d'énergie dans les bâtiments du tertiaire (vidéo).

Energiency (2013, France, 9,4M€) semble positionné sur le même créneau, mais pour les sites industriels.

[1167] Voir Artificial Intelligence Will Revolutionize Energy, Earning Billions For Investors par Ariel Cohen, 2020.

[1168] Aux USA, Honeywell propose une solution voisine. Voir Machine learning helps create "autonomous" building par Spencer Chin, février 2020.

[1169] Voir L'intelligence artificielle mise au service de l'optimisation énergétique par Chaymaa Deb, mars 2020.

App Orchid (2013, USA, $8,9M) propose une solution d'analyse visuelle de données à base de machine learning pour les utilities pour gérer leur grid et sécuriser leurs installations.

Alpiq in Tec (2009, Suisse) propose GridSense, un logiciel d'évaluation des comportements des utilisateurs pour optimiser la gestion de la consommation dans un bâtiment.

Verdigris (2011, USA, $36M) détecte la nature des appareils électriques dans un immeuble et fournit le reporting associé, une solution qui rappelle celle de **SmartImpulse** (2011, France).

Stem (USA, $351M) associe plusieurs concepts dans une offre originale : une sorte de grid de production/distribution d'énergie d'origine solaire couplée à un réseau de stockage d'énergie à base de batterie, le tout piloté à base d'IA. L'ensemble regroupe plus de 600 VPP (Virtual Power Plants), des lieux de production d'énergie solaire et une centaine de sites de stockage sur batterie. Le montant de leur financement est étonnant et très élevé.

Engie a de son côté déployé une solution d'optimisation du plan de chauffe des bâtiments en fonction de la météo, de la température extérieure et intérieure, et de la puissance de chauffage disponible. Le tout avec du deep learning associant plusieurs types de réseaux de neurones. Les bâtiments sont regroupés en clusters par type de comportements. La solution détecte les bâtiments qui dérivent dans leur classe en termes de consommation d'énergie. Cela peut conduire à des recommandations d'optimisation énergétique pour les syndics des bâtiments.

Gestion des déchets

Le tri automatisé des déchets fait maintenant appel à de l'IA à la fois côté vision, pour détecter la nature des objets et côté robotique pour leur manipulation.

Waste Robotics (2016, Canada, $2,5M) fait de la robotisation du recyclage avec un premier robot lancé en 2016 pour le tri des matières résiduelles. A base de caméras et de bras mécanique, séparation puis acheminement dans les bonnes zones de valorisation comme le compostage ou la biométhanisation. L'IA utilisée relève ici principalement de la reconnaissance d'images.

Bulk Handling Systems (1976, USA) est un autre fournisseur de robot de tri de déchets avec son Max-AI AQC-C Recycling CoBot qui est capable de réaliser un geste par seconde et de reconnaître des milliers d'objets différents (vidéo). Ils font aussi du tri de déchets de démolition. Le robot est utilisé depuis 2018 chez Véolia en France (vidéo).

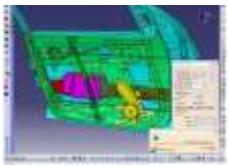
ZenRobotics (2007, Finlande, 14,4M€) est un concurrent de Waste Robotics avec son ZenRobotics Recycler (ZRR). Le bras unique du robot peut sélectionner quatre déchets de types différents en ne se trompant que dans 2% des cas, à raison de plus d'une par seconde. Il peut notamment trier des déchets de construction. Le tout fonctionne avec des caméras, des capteurs infrarouges, des capteurs de luminosité, des scanner laser 3D (comme des LiDAR), des capteurs haptiques et des détecteurs de métaux (vidéo).

sigrenEa (2009, France, acquise par Suez en 2016) intègre des capteurs de remplissage dans les conteneurs de déchets sélectifs permettant de réduire le coût de leur collecte. L'IA de planification utilise ces données ainsi que celles de la météo ou des périodes de vacances.

Industrie

L'industrie manufacturière est probablement le secteur d'activité qui exploite l'IA depuis le plus longtemps, ne serait-ce que dans les systèmes de conception assistée par ordinateurs, dans la simulation et dans la robotique de production, elle-même grosse consommatrice de vision artificielle. L'ensemble est souvent intégré dans la vague de l'industrie 4.0, un concept fourre-tout qui comprend tout ce que l'industrie a de nouveau de digital depuis quelques années, notamment les objets connectés générateurs de quantités de données utiles pour les prévisions et la maintenance, exploités par les SCADA (Systèmes de Contrôle et d'Acquisition de Données).

Les systèmes experts sont aussi couramment utilisés dans l'industrie depuis trois décennies, en particulier dans les tâches de maintenance et de réparation, même si cela fait moins l'actualité tandis que le deep learning la monopolise.



conception et simulation
CAO, PLM, VR
mise en situation
simulation d'usine

fabrication
gestion stocks et entrepots
robots de fabrication et assemblage
contrôle qualité par imagerie

exploitation
objets connectés
metering
maintenance prédictive
gestion d'assets

Nous avons déjà rapidement évoqué les robots de production dans la rubrique sur les robots.

Cette partie de cet ebook n'est cependant pas si garnie que cela avec seulement quelques pages d'exemples d'application de l'IA[1170]. Il faut dire que certaines d'entre elles sont aussi intégrées dans d'autres parties comme les précédentes sur les transports, le BTP et les utilities.

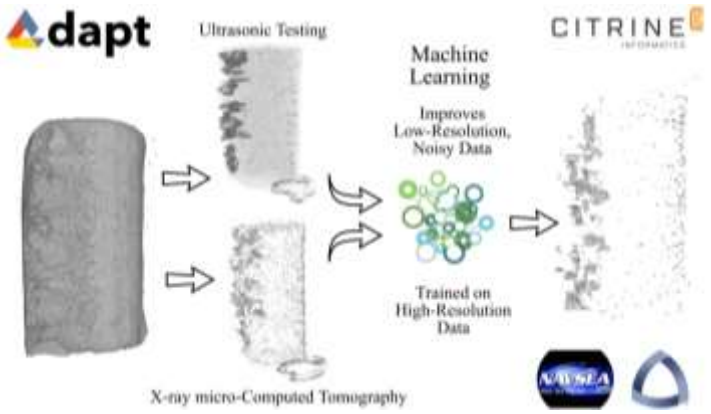
Dassault Systèmes domine le marché de la conception assistée par ordinateur avec CATIA et celui du PLM (Product Lifecycle Management) avec Enovia et est présent dans toutes les industries. Très discret d'un point de vue marketing pour ses usages de l'IA, il en utilise sans doute diverses briques un peu partout dans son offre. Les outils de simulation peuvent ainsi faire appel à des systèmes multi-agents.

[1170] De nombreuses études de cas sont présentées dans l'étude Décryptage et Cartographie Intelligence artificielle & Manufacturing, Instituts Carnot, 2020 (105 pages). Il est bien regrettable qu'il faille saisir un formulaire en indiquant son pédigree complet, adresse comprise, pour obtenir ce document.

La CAO peut exploiter des réseaux génératifs. En 2019, l'éditeur faisait l'acquisition de la startup grenobloise **Argosim** (2013, France) qui lui apporte officiellement une brique d'IA pour la conception et la modélisation de systèmes embarqués.

L'IA peut aussi servir à faire évoluer les sciences des matériaux. Ainsi, **Citrine Informatics** (2013, USA, $48M) propose une plateforme d'IA généraliste pour les industries de la chimie et des matériaux. Elle semble permettre la collecte et l'exploitation de données structurées et non structurées dans la chimie et la science des matériaux. C'est visiblement surtout un outil de R&D qui exploite le machine learning avec la visualisation de données et de la segmentation automatique de propriétés de matériaux observées expérimentalement.

L'un des principes consiste à associer les données de contrôles destructifs et de contrôles non destructifs de matériaux, ou bien de CND haute et basse résolution (*ci-contre*, source), pour identifier des corrélations entre les deux permettant de détecter des matériaux défectueux de manière non destructive.



Autre domaine d'application : la recherche de nouveaux matériaux thermoélectriques ou de verres techniques.

Leur solution a notamment permis de conçevoir des poudres d'aluminium utilisables pour de l'impression 3D métallique[1171]. Leurs principaux clients sont BASF, LANXESS (groupe chimique en Allemagne), 3M, Panasonic et les HRL Laboratories aux USA.

Un sous-ensemble de l'industrie est la prospection et l'extraction de pétrole et de gaz. Là aussi, l'IA joue un rôle émergent. Comme dans le reste de l'industrie, les processus sont surveillés avec des myriades de capteurs générant des tombereaux d'informations qui peuvent être exploités pour faire à la fois de la maintenance préventive ou prédictive et optimiser les rendements énergétiques de certaines opérations[1172].

Conception

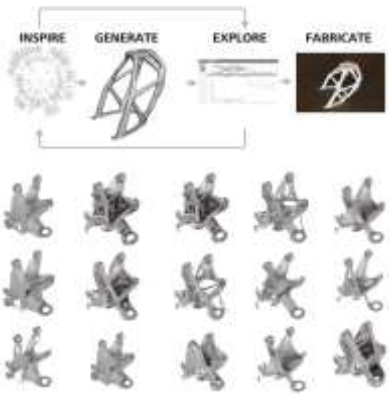
Autodesk a développé en 2017 et commercialisé en 2018 dans Autodesk Generative Design qui est intégré dans sa suite de CAO Fusion 360 une fonctionnalité de conception générative à base de deep learning génératif, ou GAN (Generative Adversarial Networks). Elle permet de générer différentes versions de designs d'objets respectant diverses contraintes. Elle exploite une base de formes et d'objets qui est automatiquement adaptée à l'objet en phase de conception.

AUTODESK

deep learning génératif pour proposer diverses formes dans la conception d'objets respectant un cahier des charge

fonctionnalité Autodesk Generative Design intégré dans l'offre commerciale Fusion 360 Ultimate

anciennement dénommé projet "Dreamcatcher"



[1171] Comme la bioinformatique, l'informatique des matériaux est un champ à lui tout seul. Voir notamment The data analytics platform for the physical world de Chris Borg (Citrine Informatics), 2008 (32 slides), Applications of machine learning and artificial intelligence to designing chemicals and materials with the desired properties de Alexander Tropsha, 2018 (64 slides), Machine Learning and Materials Informatics: Recent Applications and Prospects, 2017 (27 pages), Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence, Harvard, 2018 (109 pages) et Machine Learning and Global Optimization for Materials Discovery de Logan Ward, 2017 (52 slides).

[1172] Voir How Algorithms Are Taking Over Big Oil par Christopher Helman, janvier 2019.

Chimie et batteries

Le machine learning est aussi devenu un puissant outil permettant d'innover dans le domaine de la chimie. Il permet de déterminer la structure tridimensionnelle de molécules à partir de base d'entraînement comportant des molécules dont la structure est déjà connue.

Cela peut commencer à bas niveau par la détermination des fonctions d'ondes des électrons et donc de la forme de leur orbite[1173]. Le machine learning permet aussi de déterminer les structures de certains cristaux.

Hypergiant (2018, USA) utilise du deep learning pour optimiser le fonctionnement de son bioréacteur qui utilise des algues afin de faire de séquestrations de carbone[1174].

La conception de batteries est un domaine de la chimie qui peut aussi bénéficier du deep learning, en particulier dans la conception de la partie la plus sensible de la cathode et de ce que l'on appelle l'intercalation, là où les ions de lithium viennent s'associer à des oxydes métalliques. L'enjeu est de déterminer la bonne combinaison de ces oxydes pour minimiser l'emploi de métaux rares et/ou polluants comme le nickel et le cobalt, et d'améliorer les performances générales de la batterie (vitesse de charge, densité énergétique, puissance, maintien de la capacité dans la durée...). Un autre objectif est d'optimiser la composition de l'électrolyte où circulent les ions dans la batterie entre l'anode et la cathode[1175].

Dans ce domaine, le machine learning exploite des bases de données de composants chimiques et de tests réalisés en laboratoire. L'un de ses principaux domaines d'applications consiste à réduire le temps passé à expérimenter des combinaisons de composants chimiques. C'est une étape intermédiaire avant celle de la simulation in-silico complète d'interactions moléculaires qui relève plutôt du calcul scientifique mais qui est hors de portée pour l'instant. Entre temps, des techniques à base de machine learning sont développées pour accélérer les simulations atomiques en s'appuyant sur les équations de base de la physique quantique dans ce que l'on appelle l'*ab initio molecular dynamics*[1176].

De nombreuses approches diverses d'exploitations du machine learning sont explorées :

- Usage de **réseaux de neurones génératifs** pour évaluer des scénarios de formation d'interphase d'électrolyte solide sur les cathodes (SEI)[1177].
- Usage de **deep learning** pour déterminer la meilleure structure tridimensionnelle d'électrodes poreuses[1178].
- Recherche de **propriétés moléculaires** réalisée en 2019 par l'Argonne Lab du DoE avec un modèle entraîné avec 133 000 molécules organiques[1179].

[1173] Voir Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions par K.T. Schütt et al, novembre 2019 (11 pages). Voir aussi Artificial intelligence algorithm can learn the laws of quantum mechanics, novembre 2019, NetKet: a machine learning toolkit for many-body quantum systems par Giuseppe Carleo et al, novembre 2019 (9 pages) et Artificial intelligence algorithm can learn the laws of quantum mechanics par Kevin Kaufmann et al, janvier 2020 (35 pages).

[1174] Voir Hypergiant Is Using AI And Algae To Take on Climate Change par Sandra Ponce de Leon, janvier 2020.

[1175] Voir AI Is Throwing Battery Development Into Overdrive par Daniel Oberhaus, décembre 2020.

[1176] Voir Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning par Weile Jia et al, septembre 2020 (14 pages).

[1177] Voir A perspective on inverse design of battery interphases using multi-scale modelling, experiments and generative deep learning par Arghya Bhowmik et al, février 2019 (12 pages).

[1178] Voir Machine Learning Approaches for Designing Mesoscale Structure of Li-Ion Battery Electrodes par Yoichi Takagishi et al, août 2019 (15 pages).

[1179] Voir Building a better battery with machine learning par Jared Sagoff, novembre 2019 qui fait référence à Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations par Logan Ward et al, 2019 (13 pages).

- Couplage entre **machine learning et robots de laboratoires** avec apprentissage par renforcement pour tester rapidement des formules d'électrolytes liquides et trouver la bonne combinaison de sels en réduisant leur combinatoire[1180].
- Évaluation de la **charge de la batterie** en fonction de sa tension, de l'ampérage et de la température exploitant différentes architectures de deep learning[1181].
- Optimisation de la vitesse de charge de batteries de véhicules électriques à base de **machine learning et de modèles bayésiens**[1182].
- Prévision de la **durée de vie d'une batterie** à base de machine learning[1183]. In-fine, c'est un modèle de machine learning à base d'un simple arbre de décision qui fournit les meilleurs résultats. Des chercheurs de Stanford, du MIT et de Toyota Research Institute exploitent aussi les résultats de tests de charge/décharge en machines de tests pour déterminer ensuite le meilleur processus de charge rapide.

Du côté de l'offre, signalons la suite logicielle Alchemite d'**Intellegens** (2017, UK) qui couvre plusieurs objectifs dans la conception de batteries (recherche de matériaux, réduire l'usage de matériaux toxiques, optimiser les flux ioniques entre anodes et cathodes, développement optimum de packs de batteries et de leurs systèmes de contrôle)[1184].

On peut aussi être servi par soi-même comme **InoBat** (2018, Slovaquie) qui développe des batteries lithium-ion pour l'automobile et accélère l'identification de combinaisons chimiques prometteuses avec du machine learning.

Fabrication et contrôle qualité

Dans toutes les usines, il est courant de faire du contrôle qualité des pièces usinées avec des caméras et des solutions de computer vision, comme chez **Foxconn** qui analyse ainsi la qualité de ses cartes électroniques en sortie de bains de soudure. C'est une méthode très courante depuis des années.

un grand classique, le contrôle qualité de production par vision artificielle

source : Nvidia

Une bonne part des solutions à base d'IA dans la production servent au contrôle qualité ou à sa robotisation. Dans certains cas, cela peut aller jusqu'à la surveillance du travail des ouvriers, à défaut de les remplacer[1185].

Arcure (2009, France) développe des solutions de vision artificielle matérielles et logicielles qui détectent les personnes pour des applications de sécurisation d'usines et de chantiers. Elle sécurise notamment les déplacements de robots autonomes.

[1180] Voir Autonomous Discovery of Battery Electrolytes with Robotic Experimentation and Machine Learning par Adarsh Dave et al, décembre 2020 (11 pages) ainsi que Designing a better battery with machine learning par Austin D. Sendek et al, 2017 (24 slides).

[1181] Voir Machine Learning Applied to Electrified Vehicle Battery State of Charge and State of Health Estimation: State-of-the-Art par Carlos Vidal et al, février 2020 (19 pages).

[1182] Voir Closed-loop optimization of fast-charging protocols for batteries with machine learning par Peter M. Attia et al, 2020 (22 pages).

[1183] Voir Predicting battery life with early cyclic data by machine learning par Shan Zhu et al, septembre 2019 (5 pages) ainsi que AI accurately predicts the useful life of batteries, Stanford and MIT researchers find, mars 2019.

[1184] Voir Machine Learning for Battery Optimization par Intellegens, avril 2020 (7 pages)

[1185] Voir When AI Can't Replace a Worker, It Watches Them Instead par Tom Simonite, février 2020.

Scortex (2015, France, 4,6M€) conçoit une solution à base de FPGA pour des applications d'inspection et de contrôle qualité à base de traitement de l'image.

Deepomatic (2014, France, $9M) utilise ses solutions de vision artificielle pour faire du contrôle qualité visuel dans la production, notamment chez Valeo et Airbus. La startup semble opérer surtout en mode projet et cibler plusieurs marchés verticaux. Mais elle propose une boîte à outil de création de système de reconnaissance d'image avec son Deepomatic Studio et pour sa mise en production avec Deepomatic Run (vidéo). Elle gère l'ingestion des données, leur labellisation puis l'entraînement de modèles de réseaux convolutifs et leur déploiement en central, dans le cloud ou en partie dans les dispositifs de surveillance.

EyeSnap (2010, France) fait de la reconnaissance d'image pour l'indusrie, plutôt sous la forme de projets comme c'est souvent le cas dans le secteur.

Anodot (2014, Israël, $62,5M) propose une solution de *real-time analytics* qui s'appuie sur du machine learning pour détecter les incidents dans l'industrie en s'appuyant sur des modèles par contraintes sur des modèles de données divers.

Les robots de **Rethink Robotics** (2008, USA, $149,5M) sont aussi intéressants. Ils sont positionnés pour le travail en usine et pour collaborer avec des ouvriers humains. C'est en fait un bras articulé avec un grand nombre de degrés de liberté (vidéo). Manque de bol, la startup qui employait 90 personnes a fermé boutique en octobre 2018, non sans avoir consommé un cash incroyable et déployé ses robots chez quelques industriels aux USA[1186]. L'Allemand **Hahn Group** a ensuite acquis les marques et brevets de Rethink Robotics, toujours en octobre 2018.

Dans un positionnement voisin, **Kinema Systems** (USA) développe des solutions de vision 3D pour les robots dans la logistique et la fabrication. Ils avaient gagné le challenge Nvidia de 2018.

Pollen Metrology (2014, France, $2,4M) est une startup de Grenoble qui utilise l'IA dans une solution logicielle qui sert à la production de matériaux innovants. Elle concerne divers marchés : les semiconducteurs, les polymères, la cosmétique et la santé. Cette startup valorise des travaux réalisés au CEA, CNRS et chez IBM. Maintenant, que fait-elle qui relève de l'IA proprement dite ? En fait, il s'agit d'une plateforme logicielle, dénommée PLATYPUS, qui gère et analyse les données issues de systèmes de caractérisation d'une chaîne de production. La caractérisation est la mesure de la qualité d'une production. C'est un terme souvent employé dans la fabrication de semiconducteurs. La solution exploite du machine learning pour analyser les données et du deep learning pour analyser les tests d'imagerie de contrôle qualité. La solution logicielle s'adapte en fonction des contraintes de qualité liées au niveau de TRL (technology readyness level) de la production. Le machine learning permet de notamment de détecter en avance de phase les potentiels défauts de fabrication et d'optimiser les rendements.

DapaProphet (2014, Afrique du Sud) a l'air d'être positionné sur un domaine voisin. Leur solution à base d'IA permet de prédire et de réduire les défauts dans une chaîne de production comprenant de nombreuses étapes. Elle exploite notamment des données de production historique. La startup cible trois marchés : la construction automobile, la sidérurgie et le traitement des minéraux.

Dans un autre registre industriel, **Odesyo** (France) fait du contrôle de qualité de l'impression couleur à base d'IA d'impression couleur. Elle exploite un scanner et du machine learning pour vérifier la fidélité des couleurs après l'impression.

[1186] Ce n'est pas la seule startup de robots à avoir récemment fermé boutique, il y a aussi Mayfield Robotics avec son robot social Kuri et Jibo, un autre robot du même genre, vu au CES 2017. Voir Why Are Robotics Companies Dying? de Ron schmelzer dans Forbes, octobre 2018.

Amiral Technologies (2018, France) est spécialisé dans la maintenance prédictive industrielle. Exploitant les données de capteurs et bases labellisées, leur solution sert à prédire des défauts de vieillissement et à estimer la fin de vie utile d'un équipement industriel. Cela s'applique aux machines de production et aux moyens de transport. La startup valorise les travaux de recherche du GIPSA Lab (CNRS - Grenoble INP - Université Grenoble Alpes) via la SATT Linksium.

Petuum (2016, USA, $108M) promet dans un marketing ampoulé d'automatiser la création d'IA pour l'industrie, capable de traiter automatiquement tout le processus de création d'application démarrant par l'ingestion de données. C'est probablement très survendu. Plus prosaïquement, ils proposent BlinkLabel, un outil graphique qui simplifie la labellisation d'images qui exploite un modèle de segmentation automatique d'images, AdaptDL, un framework de deep learning auto-adaptatif et AutoDist, un framework de distribution de traitement sur plusieurs serveurs qui rappelle DeepSpeed de Microsoft[1187]. Ils opèrent dans les industries de la chimie, de l'extraction minière, des cimenteries, dans le pétrole et le gaz. Il est probable que ce genre de société propose en pratique un service outillé, à savoir la création de solutions sur mesure à partir d'une boîte à outil semi-propriétaire.

Visionairy (2018, France) propose une solution basée sur du deep learning avec apprentissage non supervisé pour détecter des défauts de fabrication dans des pièces sorties de machines outils.

Le machine learning fait même son apparition dans l'impression 3D. Il est exploitable dans différents domaines et en particulier dans la détection de défauts, dans le calibrage des imprimantes ainsi que dans les prévisions de coût et de temps d'impression[1188]. Par exemple, une équipe de recherche de l'**University of Southern California** pilotée par Qiang Huang a développé PrintFixer, un logiciel à base de réseau de neurones convolutifs qui améliore la précision des impressions et s'appuie visiblement sur une forme d'apprentissage par renforcement[1189].

Maintenance

C'est le croisement des objets connectés et de l'IA qui génère le plus de nouvelles opportunités de solutions, notamment dans la maintenance préventive et l'optimisation des ressources. La maintenance préventive des ascenseurs fait ainsi appel au machine learning chez les grands ascensoristes du marché tels que **Kone**, **Otis** et **Schindler**. Des capteurs, des bases d'entraînement, du deep learning et hop !

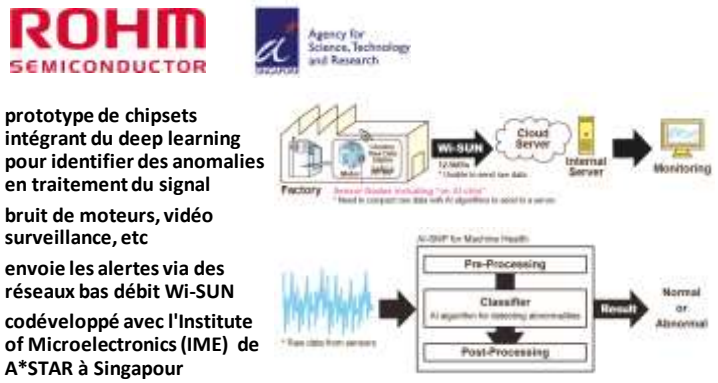
Il existe même des startups spécialisées dans le domaine comme **Uptime** (2016, France, 3M€) qui installe un boîtier dans les ascenseurs compatible avec tous les modèles du marché pour récupérer ses données de fonctionnement. Ils ont déjà quelques dizaines de clients.

[1187] Voir Une plateforme IA supervisée chez Petuum par Serge Leblal, décembre 2019.

[1188] Voir A Review of Machine Learning Applications in Additive Manufacturing par Sayyeda Saadia Razvi et al, NIST, 2019 (11 pages) et Machine Learning in Additive Manufacturing: A Review par Lingbin Meng et al, 2020 (32 pages).

[1189] Voir Shape Deviation Generator—A Convolution Framework for Learning and Predicting 3-D Printing Shape Accuracy par Qiang Huang et al, décembre 2019 (16 pages).

Le fabricant japonais de composants **Rohm** développe avec codéveloppé avec l'Institute of Microelectronics (IME) de A*STAR à Singapour des composants d'analyse du bruit de moteurs et autres signaux en exploitant du deep learning. Les alertes sont remontées en central via des réseaux de télécommunication bas débit de type Wi-SUN, un concurrent asiatique de LoRA et Sigfox, supporté cependant par STMicroelectronics.



Plus généralement, **Canvass Analytics** (2016, Canada, $5M) fournit une plateforme logicielle complète pour gérer la maintenance prédictive d'une usine. Ils le font notamment dans l'industrie automobile et dans l'agroalimentaire.

Admiral Technologies (2018, France) propose aussi une solution de maintenance prédictive, entre autres, pour les utilities et autour de la captation de données d'objets connectés divers, selon les clients. **Tellmeplus** (2011, France, $4,7M) a l'air positionné sur un créneau voisin tout comme **Prediktas** (France) qui exploite les données issues de capteurs connectés en réseaux bas débit/longue distance (LPWAN type LoRA et Sigfox ou Wi-Fi/cellulaire) et **Senseye** (2014, UK, £4,8M).

Objets connectés

Les objets connectés génèrent des données qui sont ensuite interprétées et exploitées par des logiciels à base d'IA. En voici quelques exemples.

Craft.ai (2015, France, $2M) est une startup qui était initialement spécialisée dans l'orchestration automatisée d'objets connectés avec une solution commercialisée sous la forme d'APIs destinées aux développeurs d'applications. Ils ont pivoté en 2018 et présentent maintenant leur API comme un moyen de créer des IA explicables intégrables dans des processus existants. Leur solution exploite un mix d'agents, de moteurs de règles et de machine learning. Les marchés visés sont la maintenance prédictive, la gestion des connaissances, la personnalisation d'expérience utilisateur et le CRM analytique. La startup est membre de l'association française Impact IA.

Swim.ai (2015, USA, $25,4M) développe des solutions d'analyse de données issues d'objets connectés à base de machine learning. Cela cible des usages industriels.

Ubiant (2011, France), basé à Lyon propose une solution matérielle et logicielle de gestion de la maison intelligente, de l'éclairage et de l'énergie qui s'appuie sur du machine learning et sur le Luminion, un objet connecté interagissant avec l'utilisateur via des LED de couleur indiquant si la consommation du foyer est supérieur à celle du voisinage. C'est une offre b2c.

Vivoka (2015, France, 1M€) a développé Lola, un logiciel de contrôle des équipements de la maison connectée. Elle s'appuie sur une box reliée à Internet qui se pilote via une application mobile et par commande vocale. Le projet lancé sur Kickstarter n'a pas porté ses fruits.

Ezako (2011, France) édite la solution logicielle Upalgo, qui permet d'accélérer l'analyse de données issue notamment des objets connectés couplé à Nector qui s'installe sur des kits Raspberry Pi. Ils s'appuient sur une technologie brevetée de machine learning issue du CNES qui sert à identifier les événements les plus importants dans les données pour identifier les les comportements anormaux. Ils sont déployés chez Safran, Groupe Orange.

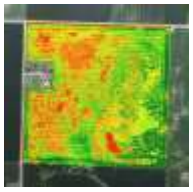
Iqspot (France, 600K€) est une startup bordelaise qui analyse la consommation énergétique des bâtiments et sensibilise ses occupants pour la diminuer. Le tout avec du machine learning. C'est une participation d'IT-Translation.

Agro-alimentaire

L'agriculture est un autre vaste domaine où l'IA a de nombreuses applications, en particulier en robotique, mais aussi en amont, avec les outils de la télédétection qui s'appuient de plus en plus sur la reconnaissance d'image à base de deep learning et sur l'agriculture de précision qui associe les objets connectés à l'IA ainsi que des robots.

Nous allons explorer ici quelques-unes de ces nombreuses innovations.

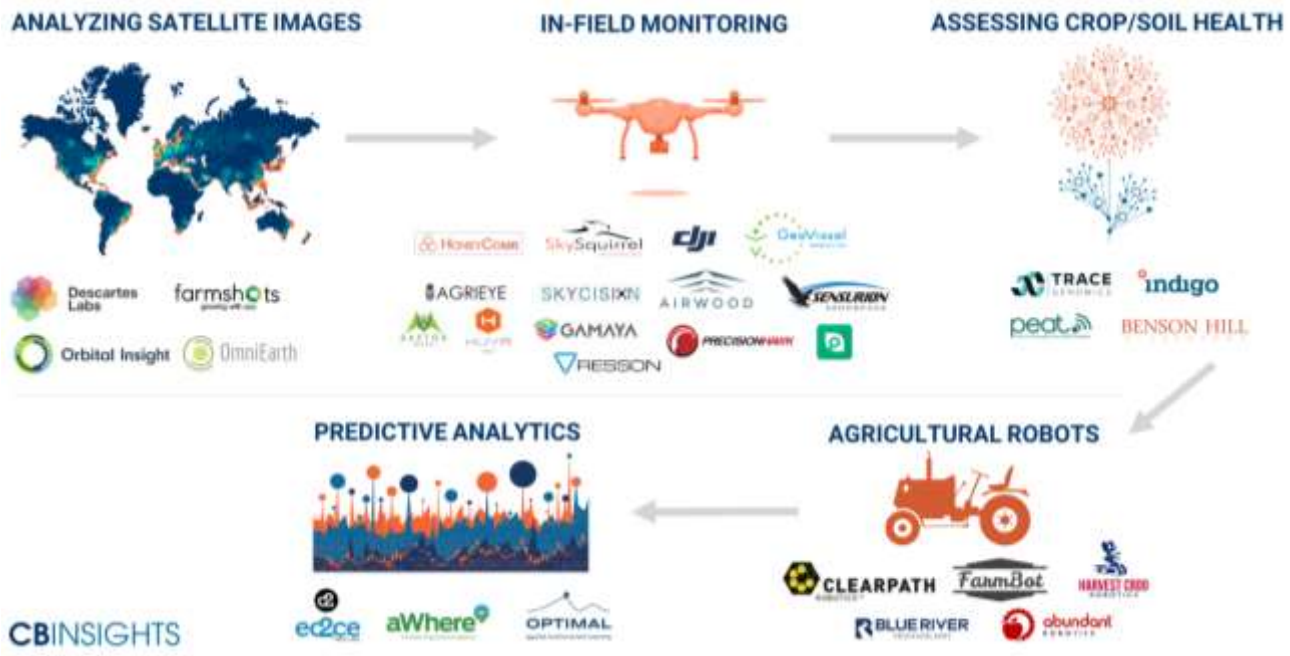
L'IA de l'agriculture utilisée tourne essentiellement autour du traitement de l'image, des données issues d'objets connectés et de la robotique.



télédétection
analyse imagerie drones et satellites
prévisions récoltes

agriculture de précision
exploitation capteurs
planification
optimisation des ressources

robotisation
binage, semis, traitements, récoltes, packaging



CBInsights[1190] décrit ce paysage avec ce beau schéma qui cartographie les startups de la télédétection de loin (satellite) ou de près (drones), de l'évaluation de l'état des champs et récoltes, des robots agricoles et des outils de prévision. Nous n'allons pas toutes les examiner ! En tout cas, il y a fort à parier qu'elles font toutes appel à des briques de l'IA.

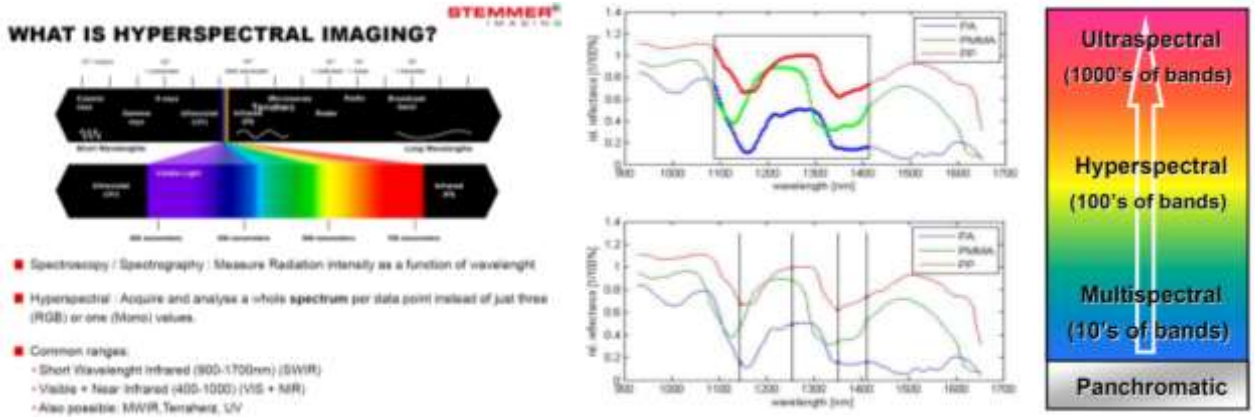
Télédétection et imagerie

L'un des premiers domaines d'application de l'IA dans l'agriculture touche à la télédétection par satellites et par drones. Il exploite le traitement d'images et notamment les variations dans le temps des observations.

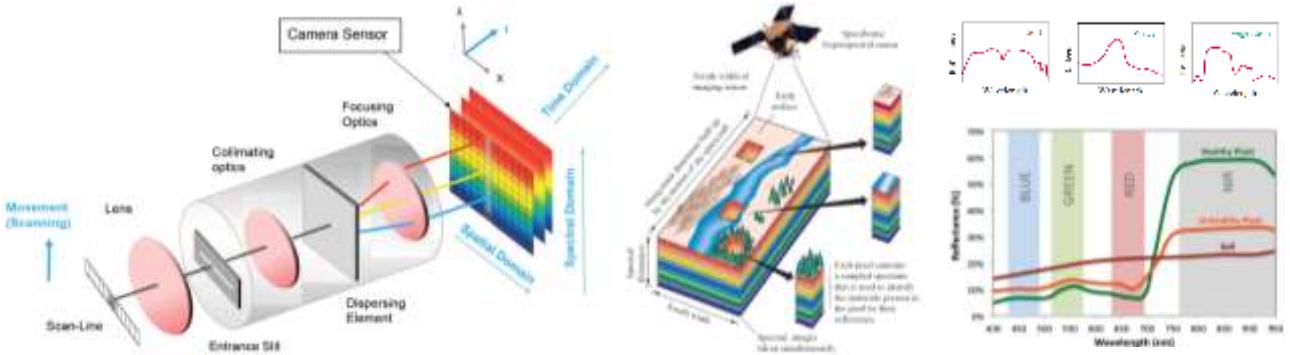
Il permet d'évaluer de nombreux paramètres comme la qualité des terrains, leur hydratation et le niveau de prévision des récoltes, à l'échelle de son exploitation aussi qu'à l'échelle globale, ce qui permet d'anticiper les cours de vente de sa production voire de les optimiser.

[1190] Source : AI, Robotics, And The Future Of Precision Agriculture, juillet 2017.

L'imagerie peut aussi servir sur terre, comme pour détecter la douleur ressentie par les animaux, ce qui peut servir à détecter des pathologies invisibles[1191].



L'imagerie aérienne profite de l'émergence et de la démocratisation des techniques de prise de vue multispectrale et hyperspectrales, qui, au lieu de se contenter de la lumière visible en rouge, vert et bleu, capte aussi les rayons infrarouges et ultra-violets pour, et va jusqu'à réaliser une spectrographie multipoints du spectre lumineux. Cela permet d'obtenir des informations très précises sur la composition des terrains, leur niveau d'engrais, les pathologies éventuelles[1192], l'hydratation, etc[1193].



En plus de **Descartes Labs** et **Orbital Insight**, déjà cités dans la partie sur la télédétection générique, nous pouvons citer également…

Farmlogs (2012, USA) qui exploite l'imagerie aérienne pour suivre l'état des champs de céréales et de coton et les traiter convenablement.

[1191] Voir Estimation of Pain in Sheep Using Computer Vision par Marwa Mahmoud & Al, 2018 (14 pages) qui porte sur la detection du niveau de douleur ressenti par un élevage de moutons !

[1192] Voir Machine learning for high-throughput field phenotyping and image processing provides insight into the association of above and below-ground traits in cassava (Manihot esculenta Crantz), juillet 2020 sur la reconnaissance de pathologies de champs de manioc par des drones.

[1193] Voir Application of multispectral imaging (MSI) to food and feed sampling and analysis, 2017 (61 pages). Sources des illustrations: Hyperspectral Imaging par Quang-Nguyen et Matthias Karl Stemmer (35 slides). Voir aussi Final report: Application of multispectral imaging (MSI) to food and feed sampling and analysis, 2017 (61 pages). L'un des fournisseurs de caméras multispectrales est MicaSense (2014, USA, $9,4M) et sa RedEdge MX qui s'embarque dans des drones comme ceux de DJI.

Farmshots (2014, USA, acquis par le Suisse Syngenta en 2018) qui détecte les maladies, et les besoins en nutriments des plantes par l'analyse d'images satellites et de drones. C'est lié à l'agriculture de précision qui vise à, par exemple, focaliser les pesticides ou l'arrosage là où ils sont le plus nécessaires. En règle générale, ces analyses utilisent un spectre électromagnétique large, comprenant notamment l'infrarouge qui détecte le mieux les variations d'états dans les plantes. **Flurosat** (2016, Australie, $2,4M) est sur le même créneau.

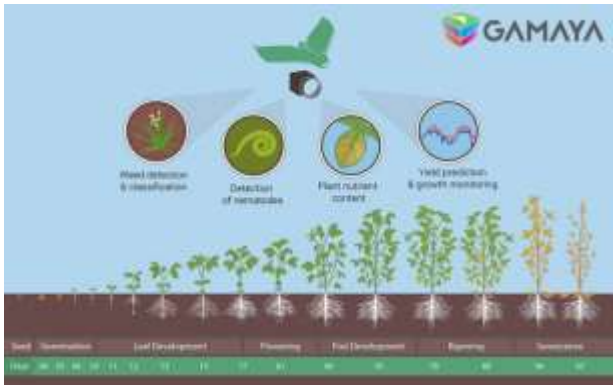
OmniEarth (2014, USA, $5,2M, acquis par EagleView Technologies en 2017) est focalisé de son côté sur l'analyse hydrique des terrains.

HoneyComb (2012, USA, $26,9M) analyse les images provenant de drones d'observation. Ils proposent le drone qui va avec, l'AgDrone System (*ci-contre*) qui provient du constructeur chinois Fly Dragon Tech.

Agrieye (2016, Ukraine, $150K) fournit comme HoneyComb une solution intégrant un drone d'observation et un service en cloud d'analyse des terrains.

Skycision (2015, USA, $1,5M) s'appuie de son côté sur les drones du Chinois DJI et les complète par ses services en cloud.

Gamaya (2015, Suisse, $4M) couvre aussi bien les grandes surfaces agricoles que les petites exploitations. Comme pour tous les acteurs de ce marché, il s'agit d'analyser les terrains pour optimiser et cibler l'usage de pesticides et d'engrais.



AirWood (2014, Inde) utilise des drones équipées de caméras multispectrales. Le reste est classique. Même chose chez **GeoVisual Analytics** (2000, USA, $1,2M).

Deepgreen (2017, USA, $500K) développe des solutions de vision artificielle pour l'équipement agricole qui permet de suivre par exemple la santé des champs.

Carbon Bee (2015, France) est un bureau d'étude en IA, électronique, opto-électronique et mécatronique spécialisé dans le développement de solutions d'imagerie agronomique pour la santé des plantes, exploitant des imageurs hyperspectraux s'intégrant dans les cultures et du deep learning. Cela permet de cibler l'épandage de produits phytosanitaires.

Agriculture de précision

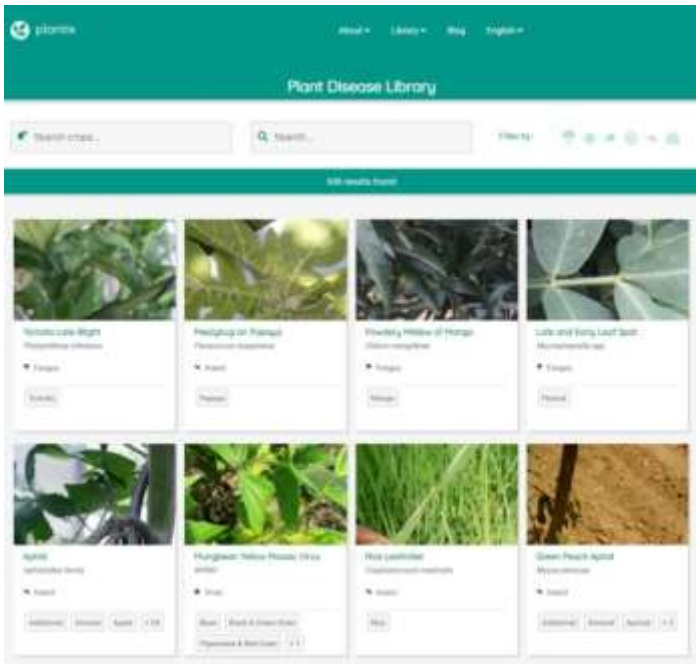
L'agriculture de précision exploite les données issues de l'imagerie aérienne et de capteurs pour assurer au mieux le rendement des récoltes, l'emploi de pesticides et d'engrais et l'arrosage[1194].

[1194] De plus en plus de startups se positionnent comme des généralistes de l'agriculture de précision. C'est par exemple le cas de **Cropx** (2013, Israël, $22,9M) et de **Semios** (2010, Canada, $17,9M).

C'est donc une affaire de consolidation de données et de machine learning permettant de détecter des anomalies et de faire des recommandations d'actions et des prévisions de récoltes[1195].

Bowery Farming (2015, USA, $167,5M) propose son système d'exploitation BoweryOS qui s'appuie sur de la vision artificielle et du machine learning pour suivre l'état de plants et optimiser leur croissance en diminuant le besoin en eau.

Peat (2015, Allemagne, $12,2M) a développé l'application mobile Plantix qui exploite de simples photos de végétaux prises par des smartphones pour identifier les maladies ou parasites qui les affectent. L'application fournit des recommandations diverses pour le traitement des plantes. L'application est gratuite et la startup espère exploiter les données récupérées, un modèle économique toujours difficile à mettre en place.



Pl@ntNet (France) est un projet lancé par des chercheurs en France qui se matérialise sous la forme d'une application mobile et web lancée en 2013 qui sert à reconnaître automatiquement plusieurs dizaines de milliers d'espèces de plantes du monde entier[1196].

Benson Hill BioSystems (2012, USA, $282,3M) a développé la plateforme CropOS servant à prédire le rendement de récoltes en fonction de différentes caractéristiques des plantes, comme leur capacité à optimiser la photosynthèse via leur ADN. La société a aussi créé un outil d'édition de gènes CRISP 2.0 censé être plus efficace que le très connu CRISP-Cas9.

FarmersEdge (2005, Canada, $103,6M) propose une solution intégrée d'agriculture de précision qui collecte les données à tous niveaux, de l'imagerie satellite aux stations météo et capteurs d'humidité implantés sur le terrain. La solution d'imagerie est capable d'identifier les évolutions des terrains dans le temps. Le tout sert à traiter les champs là où ils en ont besoin, côté engrais comme irrigation. Ce domaine comprend également **Prospera Technologies** (2014, Israël, $22M) qui utilise de l'imagerie des champs, des capteurs et des sondes météorologiques pour déterminer leur santé, le niveau de nutriments et d'hydratation ainsi qu'**Iteris** (1969, USA) qui utilise l'IA et le machine learning dans les nombreuses briques de sa solution logicielle ClearAg. Elle analyse par exemple l'humidité des sols, fait de la prévision sur le devenir des récoltes et optimise l'exploitation des données météo.

aWhere (2007, USA, $19,6M) est un spécialiste de l'intégration de données de prévisions météo pour l'agriculture et le retail. Ils font notamment des prévisions macros qui permettent d'anticiper les conditions météorologiques de pays entiers, comme pour la production de cacao au Ghana (exemple).

[1195] Voir Machine Learning in Agriculture A Review 2018 (29 pages), Deep learning for smart agriculture: Concepts, tools, applications, and opportunities, 2018 (13 pages), L'AGR-IA, une IA à la pointe du vivant (Première partie) et L'AGR-IA, une IA à la pointe du vivant (Seconde partie) par Marc Rameaux, juin 2019, Deep learning for smart agriculture: Concepts, tools, applications, and opportunities par des chercheurs chinois, 2018 (13 pages) et Deep Learning et agriculture ChaireAgroTIC, 2018 (49 pages).

[1196] Voir Artificial Intelligence for plant identification on smartphones and tablets par Hamlyn Jones, avril 2020 (7 pages) qui évalue la performance d'une dizaine d'applications mobiles grand public de reconnaissance de plantes dont PlantNet. La qualité de la reconnaissance dépend de celle de la photo et préfère des arrières plans bien neutres. Pas surprenant.

Bilberry (2016, France) identifie les mauvaises herbes à l'aide de caméras embarquées pour gérer un système de rampes de pulvérisation d'herbicide ciblé automatiques montées sur des tracteurs. Ici, le tracteur n'est pas robotisé mais le sera sûrement un jour.

Copeeks (2016, France) commercialise des capteurs divers qui permettent de suivre l'état des récoltes. Dont un capteur vidéo qui analyse l'état des champs. Il exploite le Microsoft Custom Vision SDK (vidéo).

Xarvio (Allemagne) propose Scouting, une application mobile qui détecte à partir d'une simple photo prise avec un smartphone plusieurs dizaines de maladies qui affectent les champs de blé, colza, maïs et soja à partir d'un smartphone.

Adventiel (2006, France) développe une application voisine, PhotoDiag, qui reconnaît la septoriose et la rouille brune du blé.

Chouette Vision (France) propose aux viticulteurs une solution comprenant un drone permettant le repérage et la cartographie des maladies qui affectent les vignes comme me mildiou, l'esca et la flavescence.

Plenty (2014, USA, $541M) est une ferme new wave qui cultive des salades et du kale sans OGM ni pesticides. Les cultures se font en intérieur, verticalement et sous éclairage artificiel. Le tout doit s'appuyer sur le machine learning et n'a pas besoin d'imagerie satellite ou de connaître la météo. Mais au bout du compte, ils ont remplacé un artificiel (OGM) par un autre (éclairage LED).

PowWow Energy (2013, USA, $3M) a été créé par un Français expatrié aux USA depuis 1998, Olivier Jerphagon. Leurs solutions permettent d'économiser de l'eau d'arrosage via l'usage de capteurs, de gestion de l'énergie et par l'imagerie satellite.

IBM fait aussi de l'agriculture de précision avec Watson Decision Platform for Agriculture, une plateforme qui intègre des données météo et celles qui sont issues de capteurs divers. Elle fait du prédictif pour optimiser le labourage, la plantation des cultures et les récoltes. La plateforme exploite les données de télédétection issues d'IBM PAIRS Geoscope qui sont collectées avec des satellites, drones, avions et autres capteurs (vidéo). La plateforme d'IBM est notamment adaptée aux cultures de céréales (blé, orge), de soja et de pommes de terre.

Aquabyte (USA/Norvège, $13,8M) développe un logiciel d'optimisation des coûts de la pisciculture. Il exploite la reconnaissance d'images de caméras 3D sous-marines, la température et l'oxygénation et la quantité d'aliments fournis aux poissons.

Melixa (2015, Italie) a développé une ruche connectée qui permet de suivre à distance son poids ainsi que les vols des abeilles, captés par une caméra et de l'interprétation d'imagerie.

L'application **Vectrack** développée par par **Irideon** (Espagne/Allemagne) pour l'IRTA de Catalogne (Institut de recherche et de technologie agro-alimentaire). Elle exploite des capteurs et des communications par satellite pour classifier les moustiques selon leur espèce, âge, sexe et potentiel d'infection. Des capteurs optoélectroniques sont intégrés dans des pièges qui servent à compter les moustiques et à les classifier par l'identification de leurs caractéristiques de vol. Ils sont connectés en réseau cellulaire our LPWAN (NB-IoT et LoRA) et même satellite. Le système permet par exemple de détecter la dengue[1197].

[1197] Voir L'IA pour lutter contre les moustiques par Anna Solana, janvier 2020.

Des chercheurs de l'**EPFL** ont de leur côté développé un outil de prévision de l'endroit où la foudre doit frapper. Il est entraîné avec 10 ans d'historique météo et des données de 12 stations météo en Suisse avec température, pression barométrique, humidité et vitesse du vent. Les prédictions seraient correctes à 80% ce qui est pas mal mais pas extraordinaire[1198].

J'avais sinon croisé un très grand nombre de drones agricoles chinois au CES 2018 et au CES 2019. Ceux-ci étaient surtout destinés à l'épandage sélectif de pesticides.

Robots agricoles

L'IA intervient surtout dans les techniques de robotisation d'exploitation. Comme partout ailleurs, les robots de l'agriculture sont très spécialisés. Certains s'occupent des animaux comme pour la traite des vaches mais l'essentiel est lié au cycle de vie des récoltes allant du semis aux récoltes.

Les robots présentent l'avantage théorique de permettre des économies de main d'œuvre sur des travaux qui sont en général pénibles et saisonniers. Reste à faire en sorte que cela soit rentable, les robots transférant des dépenses d'exploitation (salaires) vers des dépenses d'immobilisation (investissement dans les robots) sauf s'ils sont loués.

Nombre de robots agricoles sont surtout des projets de laboratoires de recherche qui n'ont pas pour autant abouti des années plus tard à des produits industriels. C'est le cas du projet **CASC** (Comprehensive Automation for Specialty Crops) de l'Institut de robotique de l'Université Carnegie Mellon, focalisé notamment sur les récoltes de pommes et d'autres arbres fruitiers et dont les vidéos datent de 2012. Ils planchaient même sur des robots d'estimation de taille de récolte de fruits. Le projet a été financé par le département de l'Agriculture fédéral US (USDA) à hauteur de $10M entre 2008 et 2012.

Y participait l'industriel **John Deere** qui ne semble pas avoir transformé cela en robots industriels, même s'il en exposait au CES de Las Vegas en janvier 2019, puis 2020 et même en 2021. Il est en effet encore difficile de créer des robots fiables et à des coûts raisonnables pour ces tâches.

La manipulation directe des fruits et légumes est une sacrée paire de manche. Cela fait une bonne vingtaine d'années que des robots sont mis au point pour les récoltes de fruits et légumes en tout genre : ramassage de fruits dans les arbres comme les pommes, de melons, de tomates, de fraises, de concombres[1199], d'asperges et même la récolte et la découpe de fleurs.

Ces tâches sont complexes à mener. Les robots doivent détecter les fruits et légumes de taille et formes diverses, qui sont souvent cachés derrière des feuilles ou des branches. Ils doivent ensuite les récupérer avec précaution, sans les abîmer puis les placer dans un récipient mobile.

En extérieur, les robots doivent si possible résister aux intempéries, une contrainte que l'on n'impose pas aux robots dans les usines.

Une méthode consiste à changer la forme des arbres pour les adapter à la récolte des pommes ! Un peu comme un utilisateur d'ordinateur ou de mobile s'est habitué aux idiosyncrasies fréquentes de ces appareils.

La PME française **Carré** a conçu Anatis 2 en 2014 (*ci-dessous à gauche*), un robot bineur dédié aux cultures maraîchères et équipé de caméras en tout genre pour se mouvoir et analyser le terrain (vidéo). Basées à La Roche sur Yon, la PME est malheureusement en procédure de sauvegarde d'entreprise depuis septembre 2016. Elle faisait près de 11M€ de CA en 2016, mais avec un déficit visiblement chronique de plus d'un million d'Euros.

[1198] Voir Using AI to predict where and when lightning will strike par l'EPFL, 2019.

[1199] Vous rigolez, mais un japonais, Makoto Koike, a créé un système de tri de concombres animé par une application développée avec TensorFlow (vidéo et source de l'information) ! Il utilise de la vision artificielle et détecte plusieurs paramètres : la couleur, la taille, la forme, des défauts des concombres analysés.

L'entreprise qui compte 87 salariés produisait historiquement de l'outillage agricole traditionnel. Il est difficile d'innover dans une PME traditionnelle !

Le français **Naïo Technologies** (2011, France, 3,5M€) a créé un autre robot de binage, le Oz, qui a été produit en quelques dizaines d'exemplaires (vidéo et *ci-dessus à droite*). Il permettrait de diminuer l'usage des produits phytosanitaires. Le français **SITIA** a créé un tracteur autonome (vidéo). **EcoRobotix** (2011, Suisse, $3M) a créé un robot de désherbage (vidéo) qui a un peu plus de chances d'être commercialisé que le robot **Weedmaster**, un concept du designer industriel suisse Fabian Zimmerli qui n'a pas été industrialisé.

Le robot de surveillance et de désherbage Thorvald 2 de **Saga Robotics** (2016, Norvège) permet d'éviter de faire appel à des pesticides (vidéo, *ci-dessous à droite*). Il semble qu'il s'agisse d'un projet de recherche. Il semble qu'il en soit aussi de même d'un autre robot de désherbage, le AgBot 2, de l'institut de recherche australien **QUT Research** (vidéo, *ci-dessous à gauche*).

Vitirover (2007, France) a développé un robot de désherbage des vignobles et autres types de terrain qui mesure aussi la température et l'hygrométrie au sol et détecte diverses pathologies. Il est en plus alimenté par énergie solaire.

Le laboratoire parisien **Sony CSL** a développé Lettuce Think, un robot de désherbage pour la culture de salade. Il est créé en mode open source avec un matériel léger comme un kit Raspberry Pi 2 et la réutilisation de composants existants comme une machine-outil à commande numérique 2 axes et des moteurs électriques de fauteuils roulants. Le projet a été financé par le programme européen ROMI.

BlueRiver Technology (2011, USA, $30,3M, acquis par John Deere en 2017) conçoit un robot d'épandage sélectif de pesticide (vidéo). Mais là encore, il n'a pas l'air d'être commercialisé.

EcoRobotix (2011, Suisse, $13M) commercialise un robot de pulvérisation ciblée de désherbant pour éliminer les mauvaises herbes (vidéo). Sa batterie est rechargeable par énergie solaire, ce qui augmente son autonomie.

Deux bras articulés déposent des micro-doses d'herbicide sur les zones détectées par la caméra embarquée. Il navigue grâce à un GPS et en suivant les lignes de culture dans le champ. Mais sa constitution a l'air bien fragile.

Début 2019, une quinzaine de ces robots avaient été testés, dont certains en France. Il cible notamment la production de betteraves et de colza.

CNH Industrial a créé un concept de tracteur autonome pour la récolte de céréales (vidéo) qui travaille la nuit sans broncher. Il est pilotage par tablette pour la programmation des tâches et des terrains à traiter.

Nombre de robots de récole de fruits sont ainsi encore généralement du domaine de l'expérimentation, comme pour les concombres (vidéo), les pommes destinées à la production de cidre chez SFM Technology (vidéo) ou les poivrons (vidéo).

Une startup de San Diego, **Vision Robotics** (1999, USA), s'attaque à la récolte de raisins, histoire de remplacer les travailleurs immigrés du Mexique qu'il est plus difficile de faire traverser la frontière (vidéo qui date de 2012) !

La spin-off du laboratoire de recherche SRI de Menlo Park **Abundant Robotics** (1016, USA, $12M) a créé un robot expérimental de récolte de pommes dans l'Etat de Washington, l'un des plus gros producteurs de pommes aux USA (vidéo), avec la Pennsylvanie (« Johnny Appleseed »). Il les aspire avec une sorte de ventouse (*ci-dessous à droite*)[1200].

La culture et la récolte des champignons fait l'objet de recherches à l'**Université de Warwick** au Royaume Uni. Leur robot expérimental de récolte sait détecter les champignons arrivés à maturité pour les récupérer, via l'analyse de leur taille par reconnaissance d'images. Ces robots doivent travailler 24h sur 24 et dans des environnements souterrains pas très hospitaliers.

Au Japon, les chercheurs de l'**Université d'Okayama** planchent depuis plus de 25 ans sur les robots de récolte de tomates, concombres, raisins et fraises. Une analyse spectracle dans le proche infrarouge des fruits est réalisée pour détecter ceux qui sont mûrs (sorte de Scio à usage professionnel). **Panasonic** a aussi développé son propre robot de récolte de tomates fin 2017 mais il a encore tout l'air d'être un simple prototype (vidéo). Il doit récolter 10 tomates à la minute, soit le rythme d'un ouvrier agricole mais pas mieux.

[1200] Cette intéressante présentation décrit différentes méthodes de récolte de pommes : Agricultural Robotics: Opportunities, Challenges and Perspectives de Qin Zhang, 2015 (23 slides).

Four Growers (2017, USA, $120K) conçoit aussi des robots de récolte de tomates et de raisins dans les serres mais la jeune startup est encore loin de la commercialisation. **Harvest Croo** (2013, USA, $3M) met au point un robot de récolte de fraises, qui n'est donc lui aussi visiblement pas encore commercialisé (vidéo).

On peut ajouter à cet inventaire à la Prévert le **Prospero**, un robot de semis à cinq pattes qui est toujours un prototype (*ci-dessus à gauche*).

Le robot ramasseur de fraises SW 6010 de l'Espagnol **Agrobot** (vidéo 1, vidéo 2) a l'air opérationnel, avec cueillette robotisée mais intervention humaine pour remplir les barquettes dans sa première version et entièrement automatique dans la seconde (*ci-dessous à gauche*). **Traptic** (2016, USA, $3,4M) est sur le même créneau avec un robot qui n'a pas l'air d'être encore commercialisé. L'**Université de Plymouth** (UK) a créé Robocrop, un prototype qui se focalise sur la cueillette plus délicate des framboises. Il était testé en mai 2019 (vidéo).

L'américain **Blue River Technologies** (2011, USA, $30,3M) propose de son côté un système robotisé de culture de laitues contenant une bardée de capteurs, dont certains sont 3D, pour optimiser l'entretien de laitues ou de plants de maïs (*ci-dessous à gauche*). Il a aussi lancé des tracteurs de semis de coton dont seul l'outillage est robotisé, pas le pilotage, et avec des caméras qui détectent les mauvaises herbes et un système qui dépose dessus de l'herbicide (*ci-dessous à droite*). La startup a été acquise par John Deere en 2017 pour $305M.

Les tracteurs autonomes sont plus faciles à mettre au point car ils reprennent des techniques relativement éprouvées de véhicules autonomes. C'est le cas du Bonirob des allemands **Bosch** et **Amazonen-Werke** qui est un tracteur autonome modulable (*ci-dessous*) et peut servir notamment au binage de la terre. Mais il n'a pas l'air d'être encore commercialisé à grande échelle pour autant. Le Japonais **Kubota** s'est aussi lancé dans l'autonomisation de tracteurs multifonctions[1201].

[1201] Voir Après la voiture autonome, le tracteur autonome intelligent de Kubota par Bruno Clairet, février 2020.

L'Américain **ATC** (Autonomous Tractor Company) se positionne comme le Tesla des tracteurs en les électrifiant entièrement. Ils ont aussi conçu un tracteur autonome électrique… à l'état de concept (*ci-dessus à gauche*).

L'Américain **Harvest Automation** (2009, USA, $33,5M) a créé un petit robot, le HV-100 qui déplace des pots, une tâche pas trop complexe voisine de celle des robots d'entrepôts. On en revient aux choses simples !

aGreenculture (2016, France) développe son robot polyvalent CEOL. Pesant 500 kg, il peut embarquer différents outils selon les besoins et dispose d'une autonomie de 24h. Il permet par exemple le désherbage pour la culture du maïs et le binage de cultures de lin. La startup le présente comme un « dataculteur », insistant sur les données collectées et sur la coordination entre robots du même genre.

Tibot (2016, France) développe Spoutnic, un robot avicole qui sert de rabatteur et encourage les poules à pondre au bon endroit, pour éviter le rebus lors du ramassage des œufs (vidéo). Un second robot dérivé du premier, Spoutnic NAV, circule au sein de l'élevage selon un parcours autogénéré ou défini par l'utilisateur et sert à entretenir la litière avec un outil évolutif et adaptable à l'état du sol.

Vitibot (2016, France) propose Bakus, un robot viticole polyvalent qui opère sur les pentes grâce à ses quatre roues motrices et avec une autonomie de 10 heures. Il est doté de quatre caméras « time of flight » opérant dans l'infrarouge permettant, un peu comme un LiDAR, de reconstituer l'environnement 3D (vidéo). Il est opérationnel en Champagne depuis 2019.

A terme, on verra apparaître des fermes où l'ensemble des processus sont robotisés, surtout pour les cultures sous serres. Dans la nature, les robots doivent composer avec des terrains par toujours réguliers.

Après ces quelques recherches de robots agricoles, je me rends compte finalement que l'on en est à peu près au même stade que pour les robots humanoïdes : les démonstrations et effets d'annonce sont nombreux, mais les réalisations concrètes opérationnelles le sont bien moins. Cela ne veut évidemment pas dire que cela ne marchera jamais mais que la mise au point de ces robots agricoles dans des conditions économiques satisfaisantes est encore un long chemin semé d'embûches.

En amont des récoltes, l'IA peut jouer un rôle dans le contrôle qualité, en particulier via de l'inspection visuelle. La société Suisse **Buhler** intègre des capteurs d'images dans ses systèmes d'inspection de céréales. Ils exploitent notamment la partie proche UV et proche infrarouge du spectre électromagnétique dans leur machine Lumovision et détectent les défauts des grains pouvant révéler des pathologies diverses comme l'aflatoxine, un champignon toxique qui peut se développe sur les grains récoltés. Le client peut éliminer des lots complets ou sélectionner automatiquement les grains à éliminer.

Du côté du bétail, **Alb Innovation** (2015, France) propose un robot pousse-fourrage qui rappelle les robots tondeuses et **Tibot Technologies** (2016, France, 3M€) sert à faire bouger les poulets pour leur faire prendre du poids plus rapidement.

EIO Diagnostics (2017, Canada, $720K) propose une solution de diagnostic visuel d'inflammation de pis de vache exploitant un capteur d'image multispectral fonctionnant notamment dans l'infrarouge.

Cainthus (2015, Irlande) fait de la reconnaissance de visage de vaches pour suivre leur comportement alimentaire, leur hydration, leurs mouvements et leur santé. La solution peut même identifier individuellement les vaches.

À noter enfin que l'Union Européenne a lancé le programme collaboratif **agROBOfood** de développement d'un écosystème de robotique agricole. Il est doté d'un budget de 16M€, ce qui n'est pas grand-chose[1202].

Alimentaire

L'IA peut être utilisée dans plein de registres dans les industries agro-alimentaires. Elles réexploitent des techniques vues dans la rubrique sur l'industrie concernant la fabrication de produits en usine, ne serait-ce que pour la robotisation et le contrôle qualité. C'est ce que propose **Dataswati** (2016, France), une startup qui utilise une IA multimodale exploitant de l'imagerie et des données de capteurs divers (humidité, …) pour faire du contrôle qualité en particulier dans les chaînes de production agro-alimentaires. Ils démontraient aux Universités d'été du MEDEF 2018 leur système pour évaluer la qualité de différentes tranches de saucisson ! Voilà du concret !

Quelques autres applications spécifiques plus ou moins exotiques voient le jour.

Yummly (2009, USA, acquis par Whirlpool en 2017) est une variante de Marmiton.org qui vous fait des propositions culinaires en fonction de vos goûts, du lieu, du moment dans la journée et de la saison. L'application a analysé deux millions de recettes qui ont été traitées par une IA non précisée[1203].

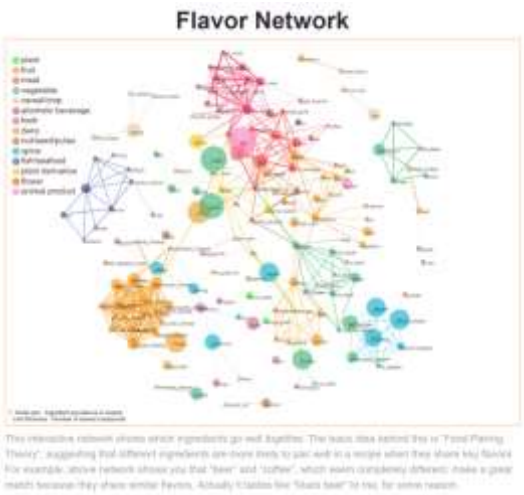
[1202] Voir agROBOfood: Towards a European network and effective adoption of robotics technologies, juin 2019.

[1203] Ces trois exemples proviennent de L'intelligence artificielle s'invite dans nos assiettes, par Morgane Thual, mars 2017.

En Israël, des chercheurs ont de leur côté créé un réseau génératif qui génère les photos de plats en lisant simplement leur recette. Mais comme de nombreux GANs, cela ne fonctionne pas encore très bien[1204].

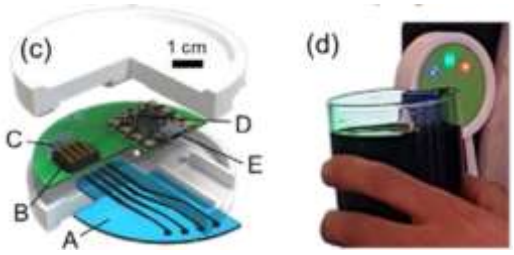
Nuritas (2014, Irlande, $61,5M) détecte les peptides actifs dans les aliments et permet de choisir ces derniers en fonction des pathologies de l'utilisateur comme certaines formes de diabète ou de pré-diabète ou d'infections. Un peptide est une petite molécule intégrant une suite d'acides aminés, comme les protéines. La société utilise un appareillage de laboratoire pour ses travaux.

Un certain Yoshiki Ishikawa a réalisé une cartographie du goût et des compatibilités entre ingrédients réalisée[1205] (*ci-contre*). **Foodpairing** (Belgique) tire parti de l'analyse de 2000 ingrédients courants et de ses 8000 molécules d'arômes pour également savoir lesquelles pouvaient être associées et satisfaire nos palais délicats. Les clients sont les industries agro-alimentaires et les chefs cuisiniers. Ce sont peut-être des algorithmes de ce genre qui sont utilisés par le brasseur **Carlsberg** qui ambitionne d'utiliser l'IA pour découvrir de nouvelles saveurs de bière[1206]. En attendant, ils pourraient recréer la Hardy's Ale écossaise qui a disparu du circuit il y a une dizaine d'années ! Pour réapparaître en 2019 !



Un autre projet de chercheurs va dans le même sens, le **Gastrograph**[1207]. Le modèle est applicable à l'envie, comme pour créer des whiskies sur mesure comme l'a récemment fait le Suédois Mackmyra avec l'aide de Microsoft[1208].

Une équipe d'**IBM** Research présentait en 2019 leur HyperState, une « langue artificielle » capable de reconnaître le goût de liquides. C'est un assemblage d'une application mobile et d'un ensemble de trois capteurs électrochimiques qui mesurent trois données: l'ionisation au chlore, au sodium et au potassium. Le machine learning fait le reste pour identifier la corrélation entre leur dosage et une base de données de liquides labellisés[1209].



Cette rubrique ne serait pas complète sans faire un tour du côté des robots cuisiniers. Les robots ménagers grand public qui rappent les carottes n'ont rien de robots. Il faut chercher plus loin pour trouver des machines qui ressemblent de plus près à des robots. On en trouve qui font le café, mais ce sont surtout des automates.

[1204] Voir This AI generates images of food just by reading the recipe, janvier 2019 qui fait référence à GILT: Generating Images from Long Text par Ori Bar El & Al de l'Université de Tel Aviv, janvier 2019 (8 pages).

[1205] Voir Flavor Network et Flavor network and the principles of food pairing, 2011 ainsi que Food-bridging: a new network construction to unveil the principles of cooking, 2017, et Data Science in Cooking de Stephanie Dodson, 2017 (20 slides), qui fait du clustering de styles régionaux de cuisine.

[1206] Vu dans Une intelligence artificielle pour aider à la création de nouvelles bières de Pierrick Labbé, juillet 2018. Les objets connectés et l'IA peuvent aussi server à perfectionner la production de la bière elle-même, comme l'a mis en place Sugar Creek Brewing aux USA, avec l'aide d'IBM. Voir Brewery Uses AI And IoT Technology To Improve The Quality Of Beer par Lana Bandoim, juillet 2019.

[1207] Voir The AI That Knows Exactly What You Want to Eat par Amos Zeeberg, 2018. Can an app lead to better-tasting food by digitally measuring flavor?

[1208] Voir AI-Created Whisky Now Available For Pre-Order par Adnan Farooqui, juillet 2019.

[1209] Voir Hypertaste: An AI-assisted e-tongue for fast and portable fingerprinting of complex liquids, juillet 2019 et A portable potentiometric electronic tongue leveraging smartphone and cloud platforms, de Patrick Ruch & Al, 2019.

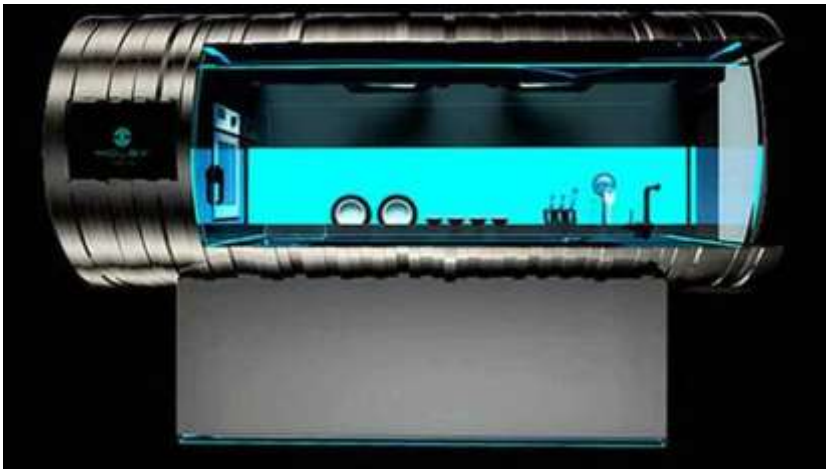
Commençons avec **Ekim** (2012, France, 12,2M€) avec son robot qui cuisine des pizzas, le Pazzi.

Il comprend deux bras articulés 6 axes (vidéo) est s'intègre dans une cuisine complète pour gérer le processus de préparation et de cuisson de la pizza. C'est compliqué à mettre au point mais cela relève de l'état de l'art, sans le faire véritablement progresser. La pizzeria autonome sera à 500K€, ce qui nécessitera d'avoir un bon flux de clients pour rentabiliser l'investissement.

Des chercheurs du MIT ont même conçu un GAN (réseau de neurones génératifs) qui créé une recette de pizza en exploitant une simple photo d'une pizza existante. Cela va sûrement changer le monde [1210]! **Domino's Pizza** utiliserait de l'IA pour vérifier la qualité de ses pizzas. C'est à base d'imagerie du produit fini et de réseaux convolutifs. La solution a été développée en Australie par **Dragontail Systems** (2013, USA).

Et puis voici une étonnante startup, **Zume Pizza** (2015, USA, $423M) qui est au départ un service de foodtrucks de pizzas californien. Ses foodtrucks sont alimentés en ingrédients comestibles et en IA pour prédire la demande des consommateurs et automatiser une partie de la production des pizzas alors que le camion est en route (vidéo). Dingue ! Mais au vu des vidéos, il semble que les pizzas ne soient pas préparées dans les camions.

En 2020, Zume se séparait de 80% de ses effectifs, soient 400 collaborateurs. Un effet du covid-19 ?



Le projet le plus intéressant est celui de **Moley Robotics** (2015, UK, $1,2M), un robot conçu par Mark Oleynik dont le lancement était prévu en 2018 puis 2019[1211]. Le prototype a été développé avec **Shadow Robots** (1987, UK). Il comprend deux bras articulés capables de manipuler les principaux outils de la cuisine, les gestes ayant été appris au robot en captant ceux d'un chef, Tim Anderson (vidéo).

Vivid Robotics (2016, USA) lançait **Picnic.** une solution pompeusement labellisée de Robotics-as-a-Service (RaaS) qui prépare des pizzas. Le robot se contente de déposer des ingrédients délivrés par de petits tapis roulants (sauce tomate, fromage, jambon...). L'ensemble est peut-être piloté par des caméras et de la vision artificielle.

[1210] Voir How to make a pizza: Learning a compositional layer-based GAN model, par Dim P. Papadopoulos & Al du MIT, 2018.

[1211] Comme cela arrive souvent dans ce marché, il est bien possible que ce robot ne soit jamais commercialisé, faute de financements et aussi, en raison de difficultés de mise au point pour le rendre véritablement opérationnel.

Il a été sélectionné par Centerplate, un traiteur dans l'événementiel qui le démontrait au CES 2020 de Las Vegas. C'est assez voisin des automates qui sont probablement déjà employés dans les usines de production de pizzas surgelées en série (vidéo).

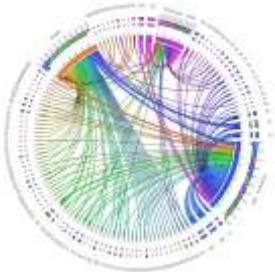
Kiwibot (USA) développe des robots livreurs de pizzas mais avec seulement 200 m de portée. Ces robots doivent être transportés par des étudiants en vélo et ils sont en fait télécommandés par des ouvriers sédentaires de la livraison sous-payés. Donc, pas tant d'IA que cela au programme !

Enfin, **Soft Robotics** (2013, USA, $25M) propose des robots de déplacement d'objets délicats comme ceux qui sont issus de la production agro-alimentaire. Ils utilisent des bras articulés avec préhenseurs en plastique ou silicone, permettant de déplacer les objets avec précaution. C'est adapté à la robotique de packaging.

Santé

L'IA est largement utilisée dans le large secteur de la santé et dans trois principaux domaines : celui du diagnostic avec diverses *medtechs* dont l'imagerie médicale ou la génomique et la médecine prédictive, celui des thérapies avec les outils pour les biotechs ou les robots chirurgiens, et enfin, sur la gestion des systèmes de santé et des soins dans la durée. La santé est donc de facto l'un des marchés verticaux les plus attirants pour les startups de l'IA avec celui de la finance et du commerce. C'est d'ailleurs un marché où les startups françaises sont assez nombreuses[1212].

diagnostics
imagerie médicale
ECG, EEG
médecine prédictive
génomique
télémédecine



thérapies
drug discovery
criblage de molécules
simulation du vivant
radio/chimio ciblée
robots chirurgiens

systèmes de santé
suivi post-opératoire
prévention des erreurs
réduction des risques
prévision de coûts

Pandémie covid-19

Il a été de bon ton de dénoncer le fait que l'IA n'avait pas pu devenir la pierre philosophale qui allait nous sauver du covid-19. Elle ne l'a ni anticipé ni prévu ce qui s'est passé pendant l'année. Elle n'a pas fait des miracles[1213]. Qu'il n'y ait pas eu de miracles est tout à fait normal : l'IA n'en fait jamais.

[1212] Voir Panorama des startups santé françaises utilisant l'IA par Mathilde Nême de Bpifrance, juillet 2020.

[1213] Voir par exemple Artificial intelligence vs COVID-19: limitations, constraints and pitfalls par Wim Naudé, avril 2020, une enquête réalisé au début de la pandémie qui montrait que l'IA n'avait pas rendu beaucoup de services, essentiellement faute de données d'entraînement. Puis Le Covid-19 a aussi fait bugger l'Intelligence Artificielle… par Benoît Zante, mai 2020, Mais où est passée l'intelligence artificielle ? par Benoît Raphael, mai 2020 et L'intelligence artificielle malmenée par la Covid-19 par Pierre Berlemont, novembre 2020.

Elle a cependant été mise en œuvre dans au moins quelques domaines, là où les données d'entraînement disponibles le permettaient. D'autres sources d'information mesurées permettaient d'en découvrir l'inventaire[1214].

La première startup à s'être fait remarquer est **BlueDot** (2008, Canada, $9,5M) qui aurait détecté l'émergence de l'épidémie en Chine. C'était un peu survendu. Leur outil passe en revue par analyse du langage des centaines de milliers d'articles de presse chaque jour et des données du trafic aérien afin de détecter et suivre les risques de propagation de maladies infectieuses. BlueDot avait prévenu ses clients le 31 décembre 2019 de l'émergence d'une épidémie en Chine, quelques jours avant que l'information devienne publique. Elle prédisait même dans quels pays le virus risquait de se propager[1215]. Leur solution permet de suivre plus de 150 types de maladies. Elle exploite des données climatiques et autres paramètres qui peuvent influencer la propagation de maladies (insectes, animaux...).

Pour le covid-19, leur système avait juste détecté un article écrit en chinois qui évoquait l'émergence de cas de pneumonie liés à un marché d'animaux à Wuhan et dont la nature n'était pas encore identifiée.

Ont suivi l'émergence d'outils d'**accélération du diagnostic** exploitant des données biologiques et issues de l'imagerie médicale.

Une équipe du **MIT** développait au début 2020 un outil logiciel de détection de cas asymptomatiques covid-19 via l'analyse de la toux par téléphone ou avec son laptop. Il n'avait cependant visiblement pas été déployé dans une application grand public. La base d'entraînement utilisait 70 000 enregistrements de toux dont celles de 2500 personnes atteintes du covid-19 à la fois asymptomatiques et symptomatiques. Le « toux » avec moins de 2% d'erreurs dans la détection[1216].

En France, **ScanCovIA** est un projet lancé en mars 2020 par l'Institut Gustave Roussy et l'hôpital Bicêtre de l'AP-HP avec la participation de l'Inria et de la startup Owkin dont l'objectif était de prédire la sévérité d'une l'infection au Covid-19 et les complications associées grâce à l'exploitation de 14 paramètres cliniques (comorbidités), biologiques et radiologiques et d'une IA. Le système exploite en particulier des scans thoraciques 3D[1217]. Le projet s'est appuyé sur les ressources du supercalculateur Jean Zay du GENCI.

La filiale d'Alibaba **AliHealth** (Chine) développait en mars 2020 une solution d'analyse de radios du poumon pour diagnostiquer rapidement le covid-19, avec une précision de 96%. C'est un cas d'usage tout à fait classique dans la mesure où l'usage du deep learning dans l'analyse d'imagerie médicale est déjà éprouvé dans pas mal de pathologies. Une solution équivalente était développée simultanément au Canada[1218]. Une équipe sino-américaine développait de son côté un système faisant l'évaluation de la sévérité de la maladie chez les patients atteints avec une solution de machine learning alimentée avec une vingtaine de paramètres décrivant le patient et autant provenant d'analyse biologique du sang[1219].

[1214] Voir Artificial intelligence cooperation to support the global response to COVID-19, par Miguel Luengo-Oroz et al, juin 2020 (3 pages), AI and control of Covid-19 coronavirus, mars 2020 et Mapping the landscape of artificial intelligence applications against Covid-19 par Joseph Bullock et al, mars 2020 (14 pages).

[1215] Voir How AI Is Tracking the Coronavirus Outbreak par Will Knight, février 2020.

[1216] Voir Artificial intelligence model detects asymptomatic Covid-19 infections through cellphone-recorded coughs par Jennifer Chu, octobre 2020.

[1217] Voir Integration of clinical characteristics, lab tests and a deep learning CT scan analysis to predict severity of hospitalized COVID-19 patients par Nathalie Lassau et al, juin 2020 (25 pages).

[1218] Voir A neural network can help spot Covid-19 in chest x-rays par Will Douglas Heaven, mars 2020.

[1219] Voir Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity par Xiangao Jiang et al, mars 2020.

Ont suivi des solutions logicielles servant à optimiser la **gestion des flux** dans les services d'urgence et dans les hôpitaux. Certaines solutions émergeaient déjà auparavant qui servent à optimiser l'allocation des ressources rares, qu'il s'agisse des personnels soignants ou du matériel de diagnostic médical[1220].

En troisième lieu, se posait évidemment la question de la **découverte de nouveaux traitements et de vaccins**. La recherche de traitements a bénéficié des mécanismes de reciblage thérapeutique à base de machine learning qui sont déjà exploités en cancérologie[1221]. D'autres techniques à base de machine learning ont permis de découvrir des anticorps opérationnels contre le virus du covid[1222]. Ces travaux se sont massivement appuyés sur les ressources de calcul de supercalculateurs, que ce soit par exemple ceux du Département de l'Énergie aux USA ou ceux du GENCI en France. Ils ont notamment exploité la technique de la simulation de dynamique moléculaire qui sert à évaluer dans le temps l'évolution de la structure 3D de molécules comme une protéase du virus, à l'échelle de la microseconde[1223].

La recherche de vaccins s'est appuyée sur des techniques de machine learning pour identifier les meilleures molécules, notamment à base d'ARN, capable d'immuniser le patient contre la glycoprotéine de spicule du virus[1224]. D'autres techniques ont été testées qui relèvent de la simulation 3D moléculaire pour déterminer la structure atomique, d'abord du virus, puis de traitements potentiels.

L'IA peut aussi être à l'origine de ratés lorsqu'on lui fait trop confiance. C'est ce qui est arrivé à la fin du printemps 2020 à l'ESN française **Devoteam** qui prédisait à 98,5% de chances qu'une seconde vague du virus n'aurait pas lieu[1225].

Un gros raté lorsque l'on voit l'impact de la seconde vague qui a démarré en octobre 2020, lié au fait que leur solution de machine learning ne devait pas disposer de suffisamment de données et de cycles épidémiques à grande échelle pour faire une prévision digne de ce nom. Ils ont pêché comme ces autres sociétés qui cherchent à exploiter les bases de données existantes pour prévoir le prochain gagnant de la coupe du monde de football !

[1220] Voir This AI Startup Helps Hospitals Navigate Operational Troubles Of Covid-19 And Beyond par Katie Jennings, 2020 qui évoque la solution de Qventus (2012, USA, $45,3M).

[1221] Voir cet exemple : Des chercheurs russes trouvent les médicaments qui bloquent la duplication du coronavirus dans l'organisme par Sputnik, mars 2020.

[1222] Voir LLNL's new machine learning platform generates novel COVID-19 antibody sequences for experimental testing, mai 2020.

[1223] Voir High–Resolution Mining of SARS–CoV–2 Main Pro-tease Conformational Space: Supercomputer–DrivenUnsupervised Adaptive Sampling par Théo Jaffrelot Inizan, Jean-Philip Piquemal et al, 2020 (19 pages). Ce dernier est aussi le cofondateur de la startup Qubit Pharmaceuticals qui développe des algorithmes de simulation quantique ou hybrides quantiques/classiques de molécules.

[1224] Voir What AI Can–and Can't–Do in the Race for a Coronavirus Vaccine par Emily Waltz, septembre 2020, Toward the systematic generation of hypothetical atomic structures: Neural networks and geometric motifs par Tess Smidt, juillet 2019 (53 slides), Computational predictions of protein structures associated with COVID-19 par DeepMind, août 2020 et AI Emerges As A Major Player In The Race To Find Covid-19 Therapies And Vaccines par Brian Uzzi, juin 2020.

[1225] Voir Covid-19: l'IA confirme à 98,5% qu'une 2ème vague n'aura pas lieu cet été ou l'hiver prochain selon une modélisation réalisée par les data scientists de Devoteam par Devoteam, juin 2020.

Mais pour les rassurer, Devoteam n'est pas un cas isolé. **HedgeChatter** (2013, USA, $385K) utilise le machine learning pour faire des prédictions diverses, concernant surtout les marchés financiers, en analysant notamment le bruit dans les réseaux sociaux. L'idée leur a pris d'appliquer cela à la prévision des cas de covid-19 et des décès associés. La première boule de cristal datant de février 2020 était de 52 millions de décès. On en est pour l'instant à environ 2 millions. Encore un raté !

C'est un peu comme lorsque l'on fait un sondage pour demander aux gens non pas pour qui ils vont voter mais qui gagnera selon eux. Les réseaux sociaux et l'agrégat d'opinions ne sont pas suffisants pour générer des prévisions fiables. Sinon, en exagérant un peu, la Terre serait plate !

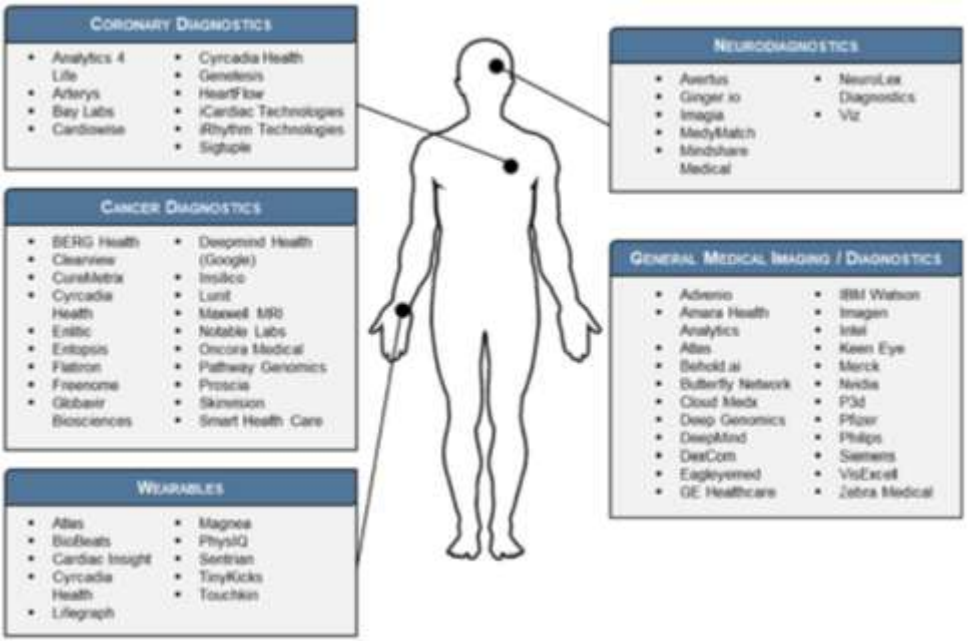
Diagnostics

L'aide au diagnostic est probablement le domaine de la santé où l'IA a le plus prospéré ces dernières années. Le deep learning et les réseaux de neurones convolutifs sont notamment omniprésents dans l'interprétation d'imagerie médicale mais pas que. Cela concerne aussi l'analyse d'examens biologiques et d'autres moyens d'investigation.

L'IA ne remplacera pas les praticiens mais les rendra plus productifs[1226]. Elle rend aussi potentiellement accessibles les connaissances des spécialistes aux médecins généralistes, ce qui est particulièrement utile dans les déserts médicaux et dans les pays émergents. L'IA permet d'augmenter la portée de certains dépistages pour éviter que des pathologies se développent silencieusement. C'est une des formes clés de la médecine prédictive, pour peu que les systèmes suivent.

Quasiment tous les pans de l'imagerie médicale sont touchés par le deep learning pour la détection de pathologies. Je vais les couvrir une par une avec quelques solutions issues de la recherche ou du marché. Il faut cependant apporter une nuance : si les prouesses de l'IA sont largement commentées dans ce domaine, les solutions associées sont encore loin d'être déployées. Les professions de santé sont naturellement conservatrices. Les bénéfices de ces solutions n'ont pas encore été évalués à grande échelle. Les régulateurs de santé sont aussi lents à la détente. Donc, il faut bien séparer les effets d'annonces de l'adoption de ces technologies.

En 2019, il y a encore assez peu de chances que vous tombiez sur un radiologue ou un cardiologue faisant appel à ce genre de solution à base d'IA.

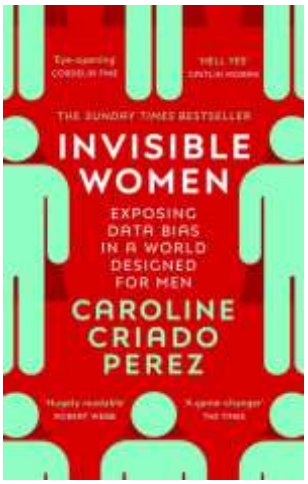


[1226] Voir AI can't replace doctors. But it can make them better par Rahul Parikh, octobre 2018 et AI Won't Replace Us, Docs Say A new survey points to a "serious rift" between the expectations of physicians and AI experts, par Megan Scudellari, décembre 2018.

Nombre d'études illustrent le fait que ces systèmes génèrent des résultats plus fiables que ceux des spécialistes[1227].

C'est un peu normal puisqu'ils consolident la connaissance collective d'un grand nombre d'experts de chaque domaine ! Ce n'est pas l'IA qui est plus forte que l'Homme. C'est un collectif d'humain qui vaut plus qu'un humain isolé, l'IA étant le médiateur de cette supériorité.

D'autres montrent que l'association d'une IA et d'un spécialiste fournit des résultats plus fiables que les spécialistes ou l'IA utilisés isolément ce qui illustre bien le fait que l'IA est un outil qui complète les spécialistes plus qu'il n'a vocation à les remplacer[1228].



Les créateurs de ces solutions à base d'IA doivent faire encore plus attention aux biais des données d'entraînement qu'ailleurs[1229]. Rien que pour ce qui concerne les femmes, ceux-ci peuvent être nombreux : pour des diagnostics, pour des prescriptions de médicaments et tout un tas de situations. Ainsi, les données d'entraînement doivent-elles tenir compte de paramètres qui affectent uniquement les femmes comme les cycles menstruels ou les évolutions temporelles de l'équilibrage hormonal. Cela peut impacter les conditions de détection de certaines pathologies, notamment dans le domaine cardio-vasculaire.

Ophtalmologie

L'IA est déjà largement utilisable pour l'analyse du fonds de l'œil afin de détecter diverses pathologies comme les rétinopathies diabétiques ou les débuts de glaucome ou de dégénérescence maculaire (DMLA). En général, les ophtalmologues dilatent la pupille des patients avec un collyre avant l'examen du fond de l'œil[1230] puis utilisent un dispositif optique qui éclaire le fond de l'œil et récupère une image. L'inspection habituelle est réalisée par l'œil de l'ophtalmologue ou par une caméra qui affiche ensuite le résultat sur un écran en temps réel. Les solutions à base de deep learning qui caractérisent les pathologies oculaires sont, à ce jour, plus performantes que les spécialistes[1231].

La détection peut concerner celle du **glaucome,** une des principales causes de cécité dans le monde qui est détectée trop tardivement dans la moitié des cas. Une expérience était menée avec **IBM Watson** en Australie en 2016.

glaucome **rétinopathie diabétique** **rétinopathie diabétique**

[1227] Voir par exemple China Focus: AI beats human doctors in neuroimaging recognition contest, juin 2018.

[1228] Voir AI-Human "Hive Mind" Diagnoses Pneumonia, septembre 2018.

[1229] Voir AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators par Carmen Niethammer, 2020 qui fait référence à Invisible Women - Exposing Data Bias in a World Designed for Men par Caroline Criado Perez, 2019 et à Data Feminism par Catherine D'Ignazio et Lauren F. Klein, 2020 (327 pages).

[1230] A terme, on pourra peut-être se passer de cette dillatation grâce à des caméras adaptées. Voir par exemple Un fond d'œil obtenu sans dilatation de l'iris grâce à une caméra bon marché, avril 2017.

[1231] Voir A.I. Equal to Experts in Detecting Eye Diseases, août 2018.

L'entraînement était basé sur l'exploitation de 88 000 fonds de rétines[1232]. La détection porte aussi sur la **rétinopathie diabétique**, qui correspond à une altération des micro-vaisseaux sanguins qui irriguent la rétine au fond de l'œil[1233]. Un premier système de ce genre a obtenu l'agrément de la FDA aux USA en avril 2018, le IDx-DR de la startup **IDx** (2010, USA, $52M).

Celui-ci avait été testé sur 900 patients avec de très bons résultats en diagnostics négatifs et positifs[1234]. C'est un dispositif médical associant matériel d'exploration optique et logiciel. Bref, c'est une solution de bout en bout[1235].

Une performance équivalente a été réalisée par des chercheurs chinois et avec un entraînement sur 80 000 fonds de l'œil et un taux de réussite de 91%[1236]. Ce genre d'examen est aussi proposé par **AIvision.healh** (2017, France) ainsi que **AiScreenings** (2017, France).

Enfin, **Eyenuk** (2010, USA, $5,9M) propose la détection de rétinopathie diabétique avec son logiciel EyeArt exploitable à partir d'un simple smartphone et d'un accessoire optique léger.

La filiale UK d'Alphabet **DeepMind** est capable de détecter une cinquantaine de pathologies oculaires d'un seul coup avec 94% de précision, en association le Moorfields Eye Hospital de Londres, dont la **dégénérescence maculaire** (DMLA) ainsi que le **décollement de la rétine**. L'outil s'appuie sur la réalisation de scans rétiniens en 3D. Il classifie les pathologies par gravité en indiquant le degré d'urgence du traitement[1237]. Le système utilise une combinaison de plusieurs réseaux de neurones convolutifs entraînés avec 14 884 scans issus de 7621 patients. Les déploiements sont prévus d'ici 2023 !

ORL

La détection des **pathologies de l'oreille** comme les oreillons et les éclatements de tympans par analyse du tympan est un autre domaine où l'IA a toute sa place pour accélérer les diagnostics[1238].

i-nside (France), fondée par Laurent Schmoll, utilise un otoscope conçu par ses soins et fabriqué par l'Allemand Stortz[1239] pour smartphone et une solution de deep learning entraînée avec 250 000 images, via le service de **Clarifai** [1240] qui peut fonctionner en mode embarqué depuis 2017 et ainsi être accessible aux praticiens dans le monde qui n'ont pas accès à Internet sur leur mobile (vidéo).

[1232] Voir Watson's detective work could help stop the silent thief of sight, février 2017.

[1233] Voir How AI Enhances & Accelerates Diabetic Retinopathy Detection, un livre blanc de **Cognizant,** février 2018 (16 pages).

[1234] La solution et l'état de l'art sont est très bien documentés dans AI is poised to revolutionize medicine. An overview of the field, with selected applications in ophthalmology de l'American Academy of Ophtalmology, 2018.

[1235] Mais les médias ont toujours tendance à généraliser. Comme ce titre : Aux États-Unis, une intelligence artificielle peut désormais faire des diagnostics médicaux, dans Top Santé, en avril 2018. Cela sous-entend que le système est généraliste alors qu'il est hyperspécialisé sur une seule pathologie.

[1236] Voir Chinese researchers develop AI technology for screening diabetic retinopathy, juin 2018.

[1237] Voir L'intelligence artificielle DeepMind, de Google, peut détecter 50 maladies oculaires aussi bien que votre ophtalmo, de Hugo Jalinière, août 2018 et L'IA DeepMind détecte les problèmes oculaires aussi bien qu'un expert humain, août 2018.

[1238] Voir Cutting Edge Technologies in Otology Field, 2017 (5 pages).

[1239] Le fondateur de i-nside, a déposé en 2013 le brevet de son Smart Scope, un petit objectif qui relie les endoscopes aux smartphones. Le brevet est utilisé sous licence par l'Allemand **K. Storz** depuis 2015.

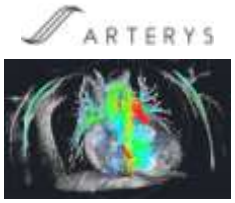
[1240] La solution d'I-nside est une étude de cas documentée par Clarifai sur https://clarifai.com/customers/i-nside.

Cardiologie

La cardiologie fait appel à plusieurs outils de diagnostic : les échographies, les radios, les IRM, les angiographies et les ECG (électro-cardiogrammes). Chacun joue son rôle dans le diagnostic ou la prévoyance. Chacun des résultats de ces analyses peut être exploité par l'IA et du deep learning.

L'analyse d'échographies à base d'IA se trouve notamment chez **Bay Labs** qui fournit aussi ses propres échographes portatifs (2013, USA, $5,5M) et **Arterys** (2011, USA, $71,7M)[1241] ainsi que chez **DIA Imaging Analysis** (2009, Israël, $12,6M).

échographies cardiaques



athéromes

ECG

Behold.ai (2015, USA, $2,2M) a développé une solution d'analyse d'imagerie médicale pour aider les radiologues à faire leur diagnostic. Le système compare les images de radiologie avec et sans pathologies pour détecter les zones à problèmes, comme les nodules et autres formes de lésions. Des chercheurs travaillent aussi sur l'analyse d'angiographies à base de deep learning[1242].

Analytics 4 Life (2012, Canada, $29M) développe son système d'imagerie cardiaque original CorVista pour la détection de pathologies coronariennes qui s'appuie sur un ECG à six électrodes fonctionnant à 8000 Hz pendant 3,5 minutes et réalisée au repos.

Il utilise une technique de « Phase Space Tomography » qui permet de reconstituer une image en 3D du cœur[1243] à partir des signaux captés et de leur déphasage, pour faire une sorte de triangulation. Le tout s'appuie sur de l'IA, mais la technique n'est pas du tout précisée dans leur littérature. Le traitement est réalisé dans le cloud mais le résultat est semble-t-il couplé à ceux d'une tomographie aux rayons X. Les études cliniques sont en cours.

HealthMyne (2013, USA, $26,4M) propose aussi un logiciel généraliste d'analyse de radios qui produit des rapports quantitatifs sur certaines observations.

Cardiologs (2014, France, $23M) est une startup française qui développe une solution logicielle en cloud d'analyse les données des ECG réalisées selon les règles de l'art avec plusieurs électrodes (4 sur les membres, 6 sur le thorax) en cabinet médical, par des infirmiers ou des docteurs. Les résultats sont fournis sur une interface web. Elle s'appuie sur des méthodes de machine learning exploitant des réseaux de neurones convolutifs avec apprentissage supervisé (ConvNets). Côté cloud, ils utilisent comme nombre de startups de l'IA les ressources de Google Tensorflow.

[1241] Arterys a aussi obtenu en février 2018 l'agrément FDA pour la commercialisation d'une solution à base d'IA de détection de cancers du foie et du poumon. Voir Arterys Receives First FDA Clearance for Broad Oncology Imaging Suite with Deep Learning, février 2018.

[1242] Voir AI Technology Can Help Speed Up Stroke Diagnosis And Treatment par Helen Albert, septembre 2020, au sujet de l'analyse automatique d'angiographies pour détecter des thromboses artérielles cardiaques. Cela reste un outil d'aide au diagnostic car il génère des faux positifs.

[1243] Le principe de la phase tomographie est expliqué dans Can Phase Space Tomography be Spun-off back into Medical Imaging? (10 pages).

Cela permet de fournir une réponse en quasi-temps réel. Ils ont entraîné leur système avec des bases d'ECG dont une base de 100 000 ECG venant du Minnesota récupérée en 2015. Il faut payer pour, mais ce n'est pas trop cher ! Leur système est capable de prédire une centaine de troubles sur 12 canaux (ECG au repos) et une quinzaine sur 1 à 2 canaux (ECG ambulatoire).

Le système détecte notamment la fibrillation atriale, qui est corrélée à l'apparition d'AVC – accidents vasculaires cérébraux – du fait d'une mauvaise circulation du sang, dont le risque augmente avec l'âge et qui est plus facile à traiter lorsqu'elle est précoce.

L'analyse d'ECG passe souvent par l'utilisation des plusieurs méthodes et différents réseaux de neurones (RNN, CNN, LSTM)[1244]. L'efficacité des méthodes dépend de la pathologie à détecter. Par exemple, la fibrillation attriale est mieux détectée avec un réseau convolutif. Cardiologs a obtenu l'agrément de la FDA en 2017 pour couvrir le marché US ce qui est assez rare pour une medtech française.

L'Apple Watch 4 annoncée en septembre 2018 et légèrement améliorée dans la version 5 en septembre 2019 intègre aussi une fonction d'électrocardiogramme. C'est un ECG à trois points qui ne détecte visiblement qu'environ 10% des pathologies cardiaques[1245]. Elle détecte aussi les anomalies de rythme cardiaque. Il faut donc éviter de la survendre et ne pas oublier qu'un ECG réalisé chez un cardiologue sera plus riche en information ! Tout dépend de la pathologie recherchée ou à surveiller ! L'Apple Watch 6 lancée en 2020 ajoutait la mesure de l'oxygénation du sang, une fonction utile en période de pandémie covid-19, mais apparemment sans rapport avec l'IA.

La prudence est autant de mise avec cette performance de **Google** et **Stanford** consistant à prédire des pathologies cardiaques par l'analyse du fond de l'œil. Réalisée en 2018, elle s'appuyait sur un jeu d'entraînement de 284 335 patients validé sur deux jeux de test de 12 026 et 999 autres patients. Le taux d'erreur semble rester élevé[1246], et les chercheurs se gardent bien de le comparer avec celui des méthodes traditionnelles. Si le taux d'erreur d'un fond de l'œil est plus élevé qu'une échographie ou une IRM, pourquoi s'embêter à faire un fond de l'œil ?

Il faut adopter les mêmes réserves face à des annonces parfois relayées de manière un peu simpliste. Ainsi, en mai 2019, on apprenait qu'une IA pouvait « **prédire la mort mieux qu'un médecin** »[1247]. Une IA entraînée avec des dizaines de données de près de 1000 patients étalées sur six ans pouvait établir des corrélations entre la mort ou les attaques cardiaques et ces différentes variables, le tout avec une précision 90%. En creusant un peu[1248], on se rend compte que la solution en question dénommée **LogitBoost** exploitait bien 85 variables de 950 patients mais ceux-ci étaient déjà suivis pour des douleurs thoraciques !

[1244] Voir Machine Learning in Electrocardiogram Diagnosis, 2009.

[1245] L'une des méthodes est décrite dans End-to-end Deep Learning from Raw Sensor Data Atrial Fibrillation Detection using Wearables 2018 (7 pages). Il s'agit ici de détecter la fibrillation atriale avec un réseau convolutionnel couplé à un LSTM (réseau à mémoire). Cette pathologie peut notamment entraîner un AVC.

[1246] Voir Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning, février 2018 (21 pages).

[1247] Voir L'intelligence artificielle peut à présent prédire la mort mieux qu'un médecin par Bastien L, mai 2019.

[1248] Voir Machine learning overtakes humans in predicting death or heart attack, mai 2019.

Il s'agissait donc de patients présentant déjà des facteurs de risques cardiaques. De plus, l'étude montrait surtout que la prévision de troubles cardiaques est nettement améliorée en intégrant les résultats d'imagerie médicale, ce qui n'a rien de très surprenant. La prévision s'appuyait sur la combinaison complexe de nombreux facteurs de risques. Le modèle n'est pas parfait. C'est surtout un outil d'aide à la décision pour les praticiens. D'autres études montrent que les prédictions concernant les patients sont assez aléatoires et doivent être prises avec des pincettes[1249].

VoltaMedical (2016, France, 2,6M€) aide les chirurgiens à anticiper les risques de fibrillation auriculaire ou atriale avant des opérations du cœur. Les données d'entraînement exploitées comprennent notamment des électrocardiogrammes labellisés.

Bitmakers (2016, France) a créé Heartbit, qui exploite les données d'électrocardiogrammes (ECG) générés avec des holters portés par des patients en suivi ambulatoire. L'application tourne sur smartphone.

Eko (USA) est un fabricant de stéthoscopes qui a développé une IA à base de réseaux de neurones qui analyse le murmure du cœur « *mieux que les cardiologues* ». Une promesse classique.

Nuvoair (2015, Suède, $5,2M) gère les insuffisances respiratoires avec un spiromètre couplé à leur application maison ResAppDx qui détecte différentes pathologies respiratoires en analysant le son de ta toux.

Archeon (France) est une startup de Besançon qui a développé EOlife, une solution à base de machine learning qui évite de générer une phase d'hyperventilation lors du processus de réanimation cardiaque destinée aux secouristes.

Enfin, citons cette expérience menée par des chercheurs américains de l'**Université de Washington** qui a permis de détecter les symptômes de crise cardiaque en analysant le souffle d'utilisateurs isolés via des enceintes connectées. L'idée est évidemment d'exploiter cela pour appeler rapidement les secours. Le taux de reconnaissance serait très bon avec moins de 1% d'erreurs.

Cancérologie

Les cancers sont la seconde source de mortalité après les maladies cardio-vasculaires, avec une incidence de coût très élevée pour la société. D'où le fait que la cancérologie est l'un des plus gros marchés de la santé. Il est fragmenté en de nombreux types de cancers différents qui ont leurs propres techniques de diagnostics, s'appuyant sur différentes formes d'imagerie médicale et d'analyses biologiques.

Presque tous les cancers ont leur solution spécifique d'analyse d'imagerie médicale à base d'IA ou d'analyse plurifactorielle. Tout y passe : le poumon, le foie, le colon, le pancréas, le sein, le cou et la tête. Mais les diagnostics du cancer passent aussi par l'analyse de biopsies, aussi par imagerie, ainsi que de résultats biologiques aussi bien de marqueurs classiques que de génotypage ou séquençage de l'ADN.

Les **cancers du poumon** sont qualifiés par analyses de radios chez **Enlitic** (2014, USA, $15M) ainsi que chez **Riverain Technologies** (2004, USA) qui est focalisé sur la cage thoracique. En 2019, Des chercheurs de **Google AI** créaient avec **Northwestern Medicine** (USA) une IA de diagnostic du cancer du poumon qui serait meilleure que les radiologues.

La différence de taux de reconnaissance avec six de ces derniers était de 9,5%, ce qui n'est pas grand-chose. Ils utilisent un réseau de neurones qui travaille en 3D au lieu de 2D[1250].

[1249] Voir Projecting the outcomes of people's lives with AI isn't so simple par Virginia Tech, mars 2020.

[1250] Voir Google's lung cancer detection AI outperforms 6 human radiologists par Khari Johnson, 2019. Pour les analyses de scans en 3D, voir End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography par Diego Artila & Al, mai 2019. et 3D Deep Learning from CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas par Zhao W, 2018.

Des chercheurs chinois ont de leur côté créé un modèle de prévision de la prolifération d'adénomes du poumon[1251]. **Infervision** (2015, Chine, $75M) fait de la détection du cancer du poumon avec des scans PET. Le produit est déjà déployé dans divers hôpitaux en Chine. D'autres méthodes permettent cette détection mais avec des images 3D[1252]. La **Chester AI** développée par Joseph Paul Cohen du Mila et de l'Université de Montréal est une solution open source qui analyse les radios du poumon (vidéo). Elle détecte 14 pathologies différentes et fonctionne dans un navigateur web. Les images restent sur l'ordinateur et ne sont pas envoyées dans le cloud[1253].

Le **cancer du foie** est aussi au programme. **Guerbet** (France, 467 M€ de CA en 2017) et IBM annonçaient en juillet 2018 le codéveloppement d'une solution dénommée Watson Imaging Care Advisor for Liver pour son diagnostic. Elle doit exploiter les résultats d'imagerie scanner et IRM et améliorer la caractérisation des métastases qui sont complexes à analyser dans ce type de tissus.

Les **cancers du cerveau** sont aussi au programme des réjouissances[1254] ! Une startup francilienne, **Qynapse** (2015, France), analyse de manière itérative les résultats d'IRM cérébrales pour suivre l'évolution de traitements, notamment dans la lutte contre les cancers du cerveau.

Les **cancers du côlon** sont détectés par analyse d'image des colonoscopies[1255]. Il existe aussi des solutions de diagnostic du **cancer du pancréas**[1256] ainsi que de **la tête et du cou**[1257].

Les **cancers du sein** sont analysés chez **Volpara Solution** (2009, Canada, $5,5M + IPO) qui réalise des analyses densitométriques précises et aussi chez **QVViewMedical** (2006, USA, $4,75M) ainsi qu'avec **Therapixel** (2013, France, $5,6M). Ces systèmes peuvent notamment être entraînés avec les 640 000 images issues de 86 000 patientes, récupérées dans la base du Digital Mammography DREAM Challenge lancé aux USA entre 2016 et 2017[1258]. Therapixel avait d'ailleurs gagné mi-2017 la première étape de ce challenge. Therapixel se distingue des autres solutions en expliquant son diagnostic. Avec sa solution Mammoscreen, il se positionne en complément des spécialistes, pas pour les remplacer. L'un de ses concurrents est Profound IA de **iCAD** (1984, USA, $37,5M)[1259]. La filiale **DeepMind** de Google montait les enchères en 2020 en publiant une étude selon laquelle leur système de détection du cancer du sein générait un très faible taux de faux positifs et de faux négatifs[1260].

Le **MIT** a créé de son côté un modèle de deep learning qui peut prédire l'apparition d'un cancer du sein jusqu'à cinq années à l'avance. Il est entraîné avec les radios de 60 000 patientes, à équité entre personnes blanches et noires. Comme les femmes noires ont plus de chances de voir apparaître un tel cancer, cela supprime des biais de faux négatifs dans le système[1261].

[1251] Voir 3D Deep Learning from CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas, 2019.

[1252] Voir S4ND: Single-Shot Single-Scale Lung Nodule Detection, juin 2018 (8 pages).

[1253] Voir This free AI reads X-rays as well as doctors, par Marc Wilson, janvier 2019.

[1254] Voir Brain Tumor Segmentation and Tractographic Feature Extraction from Structural MR Images for Overall Survival Prediction, 2018 (12 pages) appliquée à la détection précise de glioblastome, la pathologie qui a récemment emporté le sénateur américain John McCain.

[1255] Voir Artificial Intelligence Automatically Spots Polyps During Colonoscopies, 2018.

[1256] Voir Pancreatic Cancer Collective funds two AI teams to identify high-risk populations, mars 2019 ainsi que Artificial intelligence brings pancreatic cancer screening one step closer to reality, juillet 2020 qui, comme le titre l'indique, illustre le fait que ces solutions ne sont pas encore déployées.

[1257] Voir Machine learning improves the diagnosis of patients with head and neck cancers, septembre 2019.

[1258] Voir The Digital Mammography DREAM Challenge ainsi que DREAM Challenge results: Can machine learning help improve accuracy in breast cancer screening? par IBM Research, juin 2017.

[1259] Voir Profound IA, l'algorithme de détection du cancer du sein déjà en service par Coralie Lemke, uin 2019.

[1260] Voir Will DeepMind's Breast Cancer Diagnosis AI Replace Specialists? par Daniel Shapiro, 2020.

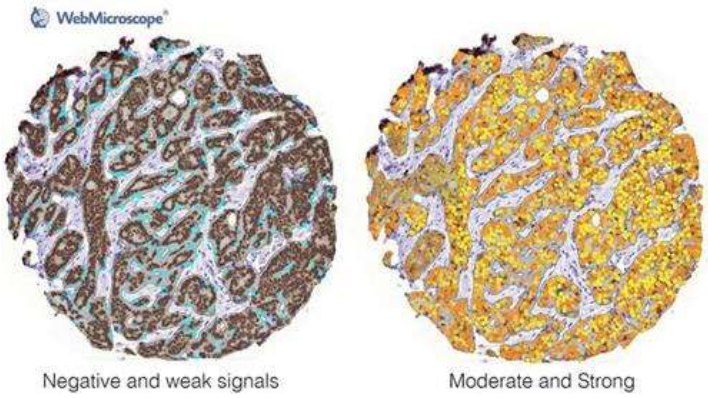
[1261] Voir Using AI to predict breast cancer and personalize care par Adam Conner-Simons et Rachel Gordon, mai 2019.

Imagia (2015, Canada) propose sa solution « Deep Radiomics » qui analyse l'évolution dans le temps de cancers par imagerie médicale couplée aux données cliniques des patients avec notamment le cancer du côlon[1262].

Le vaste champ de l'**anapathologie** est aussi à même d'exploiter les solutions d'analyse d'image à base de réseaux de neurones. L'anapat sert à analyser les cellules de biopsies, notamment dans le diagnostic de cancers aussi bien liquides que solides.

WebMicroscope (2013, Finlande, $6,5M, aussi dénommé Fimmic) réalise ses analyses dans le cloud à l'aide de GPU. Comme de nombreuses solutions d'IA en imagerie médicale, elle détecte des cellules cancéreuses et apporte surtout des indications quantitatives par comptage de cellules (*exemple ci-dessous*).

KeeLab (2018, France) est positionné sur ce même marché de l'analyse en anatomie pathologique de cellules cancéreuses mais plus au niveau qualitatif. Ils cherchent à minimiser le taux de faux positifs, en-dessous de 1% sachant que c'est un outil d'aide au diagnostic pour les « anapats ». Ils ciblent les cancers du sein, du système digestif, de la prostate et du poumon.



PathAI (2016, USA, $75,2M) fait de même[1263], tout comme **Paige AI** (2018, USA, $25M) qui a construit son propre supercalculateur à base de Nvidia V100 totalisant 10 petaflops et se focalise sur le cancer de la prostate, **Ibex Medical Analytics** (2016, Israël, $13,6M) qui s'attaque aussi au cancer de la prostate, **Deep Lens** (2017, USA, $17,2M) utilise l'imagerie pour l'anapat d'identification de patients candidats à des tests cliniques. C'est aussi l'activité de **Keen Eye** (2013, France, 7,5M€).

La détection de cellules cancéreuses s'applique aussi aux globules blancs pour les leucémies, ce que publiaient des chercheurs allemands en 2019. Leur système était entraîné avec une base labellisée de 20 000 images individuelles de globules blancs[1264].

Nucleai (2017, Israël, $11,5M) développe des solutions d'anapat servant à évaluer l'efficacité dans le temps des traitements d'immunothérapies. Le fondateur vient de l'armée israélienne et était spécialisé dans l'interprétation d'imagerie aérienne. Leur dernier tour de financement en 2020 a été réalisé avec l'entreprise de pharmacie suisse Debiopharm.

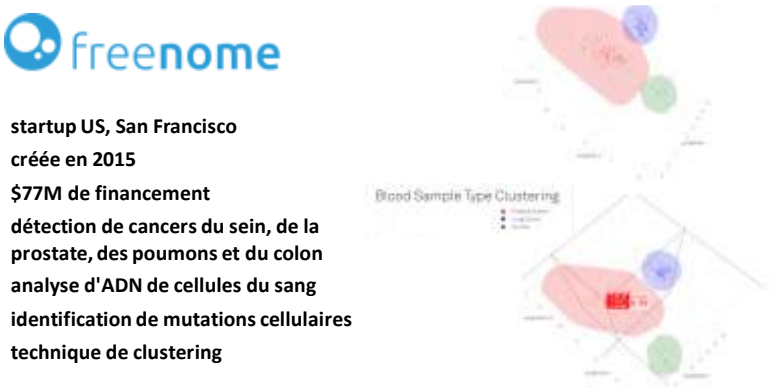
VitaDX (2015, France) détecte le **cancer de la vessie** dans les cellules des urines avec un procédé voisin.

[1262] Voir AI Innovators: Radiology Startup Puts The Power Of AI In The Hands Of Physicians par Mona Flores, 2018.

[1263] Voir The First Frontier for Medical AI Is the Pathology Lab par Elie Dolgin, novembre 2018.

[1264] Voir Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks par Christian Matek et al, novembre 2019 (7 pages).

Freenome (2015, USA, $77,6M) produit des analyses de biopsies liquides – essentiellement du sang - permettant la détection de cancers émergents, du sein, de la prostate, des poumons et du colon (vidéo). Ils s'appuient sur des analyses génotypiques et sur le clustering de résultats. Reste après à traiter les cancers découverts, ce qui est une autre paire de manche !



En juin 2019, **Adaptive Biotechnologies** (2009, USA, $406,9B) et Microsoft annonçaient être en train de mettre au point un test sanguin mâtiné d'IA permettant de détecter diverses maladies. Cela semble extraordinaire. En pratique, c'est un procédé d'analyse par séquençage génétique des globules blancs du système immunitaire (les T-cells). Donc, comme la plupart des tests sanguins, un test très ciblé. L'IA associée permet de faire la corrélation entre les analyses et l'émergence de cancers et maladies infectieuses entraînées par des désordres du système immunitaire[1265].

Le machine learning doit aussi permettre d'éviter la détection erronée de tumeurs entraînant des opérations inutiles, coûteuses et dangereuses. C'est la lutte contre les faux positifs. L'un des exemples est le cas de la vésicule biliaire, un organe très sensible rattaché au foie. Un diagnostic d'une tumeur détectée au scanner peut enclencher une opération qui va mobiliser le patient pendant trois mois avec 50% de chances de complications sans compter 5% de risque de décès sur le billard. Aux USA, ce type d'opération ne serait pas justifié dans 78% des cas, les tumeurs n'étant pas bénignes. Il en va de même pour ls kystes du pancréas.

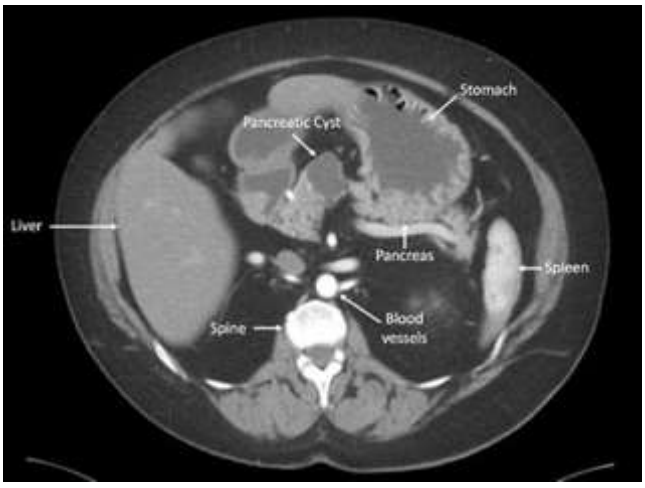
Ce même Microsoft développait en 2019 avec le laboratoire **SRL Diagnostics** une solution à base de machine learning de détection du cancer du col de l'utérus avec des IRM, visant en particulier le marché indien[1266].

La solution CompCyst de la **Johns Hopkins University** permet d'éviter cela[1267]. Elle vient de sortir de l'expérimentation. Elle exploite des données de centaines de patients opérés issues de 15 centres médicaux dans le monde. Les données comprenaient les examens des kystes du pancréas extraits par opérations.

La technique de machine learning utilisée est MOCA (Multivariate Organization of Combinatorial Alterations) qui combine des données de mutations génomiques, l'analyse de protéines et les résultats d'imagerie médicale.

Le ML a été entraîné avec les données de 436 patients, ce qui n'est pas énorme, puis testé avec un autre jeu de données de 426 patients.

En gros, CompCyst permettrait d'éviter d'envoyer inutilement au bloc opératoire entre 60% et 74% de patients, sans générer de faux négatifs.



[1265] Voir Microsoft's AI Diagnosing Blood Test Is Only A Few Years Away From Launch par Tyler Lee, juin 2019.

[1266] Voir Microsoft AI diagnoses cervical cancer faster par JonFingas, novembre 2019.

[1267] Voir With This AI, 60 Percent of Patients Who Had Surgery Could Have Avoided It par Megan Scudellari, juillet 2019.

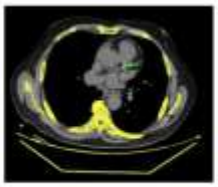
La solution devrait être commercialisée par **Thrive Earlier Detection** (2019, USA, $110M), une startup très bien financée ayant licencié la propriété intellectuelle du projet CompCyst.

Système osseux

Il fait l'objet de radios qui ne sont pas forcément difficiles à interpréter.

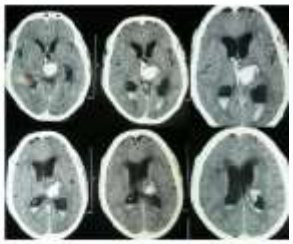
L'**ostéoporose** est pourtant bien caractérisée dans l'analyse de radios chez **Zebra Medical Vision** (2014, Israël, $50M) qui détecte aussi les compressions de vertèbres, la stéatose hépatique (foie gras) et les hémorragies cérébrales. Et ils détectent aussi le cancer du sein. Bref, ce sont des spécialistes qui deviennent des généralistes.

ostéoporose



stéatose hépatique

compression de vertèbre



hémoragies cérébrales

Gleamer (2017, France, $10M) développe une solution d'aide à l'analyse d'imagerie médicale pour les radiologues. Leur première solution est BoneView, qui dédiée à l'analyse de radios osseuses en traumatologie et y détecte les microlésions. Le système assiste les radiologues au lieu de les remplacer. Il réduit de 30% le taux des fractures non-détectées dans les radios et accélère au passage la durée de l'interprétation des radios. La startup indique que sa solution était utilisée (en septembre 2020) par une cinquantaine d'hôpitaux et de cliniques en France dont l'Hôtel Dieu et l'hôpital Ambroise Paré.

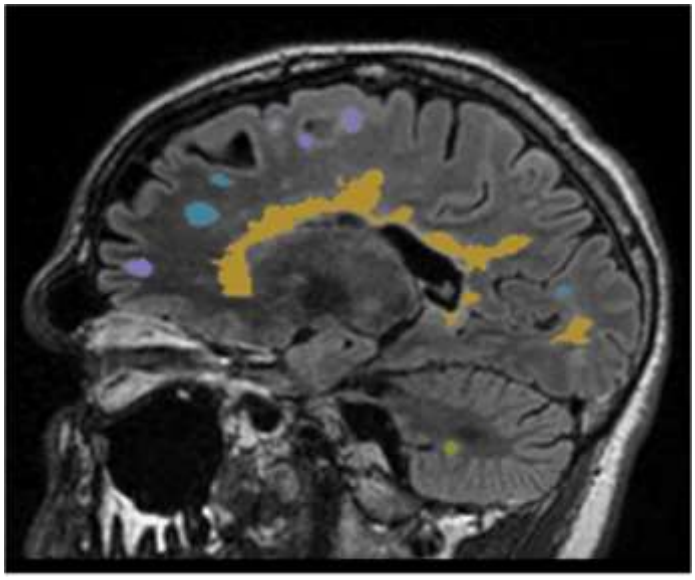
Système nerveux

L'IA peut aussi bien servir à mieux comprendre le fonctionnement du cerveau qu'à en détecter certaines pathologies et troubles cognitifs. Les outils de diagnostic à base de machine learning de corrélation ou de deep learning d'analyse d'imagerie médicale ne manquent pas. Ils sont pour l'instant à l'état de travaux de recherche et pas encore déployés à grande échelle dans les systèmes de santé. En voici quelques exemples illustratifs.

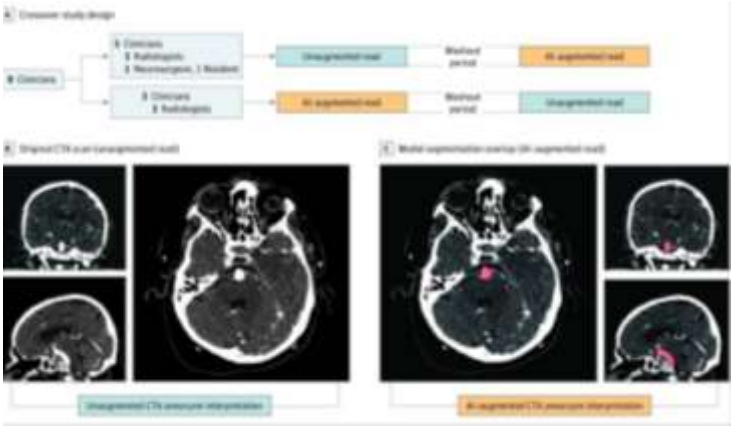
Des chercheurs de l'**Université de Chicago** découvraient en 2019 que la mémoire à court terme utiliserait des circuits neuronaux différents selon les tâches, ce grâce à des réseaux de neurones artificiels servant à simuler leur fonctionnement des neurones biologiques[1268]. Ils ont découvert un processus de mémorisation « silencieux » sans activité neuronale détectable et un processus plus actif qui, lui, y fait appel. Le tout avec des capacités mémorielles de quelques secondes à quelques minutes, témoignant de la plasticité de la mémoire à court terme.

[1268] Voir AI software reveals the inner workings of short-term memory, juin 2019 qui fait référence à Circuit mechanisms for the maintenance and manipulation of information in working memory, par Nicolas Masse & Al, 2019 (42 pages).

L'intérêt de ce type de recherche est à la fois de mieux comprendre le fonctionnement du cerveau mais aussi de trouver de meilleures méthodes pour l'imiter dans l'IA.



Les pathologies du cerveau sont détectées par le Belge **Icometrix** (2011, $2,17M) avec son logiciel Msmetrix qui analyse les résultats d'imagerie médicale pour détecter les lésions, mesurer leur volume ainsi que celui du cerveau avec des applications dédiées à la sclérose en plaques (*ci-contre*). Des chercheurs anglais ont de leur côté créé une méthode de détection de lésions de petits vaisseaux dans le cerveau[1269].



Une équipe de chercheurs de **Stanford** créait en 2019 un système de détection des anévrismes cérébraux exploitant de l'imagerie scanner à rayons X couplée à un agent de contraste sanguin. Entraîné avec seulement 611 scans labellisés, le système a permis aux radiologues de détecter 6 cas supplémentaires d'anévrismes sur 100 scans. Le test était réalisé avec 8 cliniciens évaluant 115 scans comprenant ou pas des anévrismes[1270].

Ce genre de performance est dans la lignée de celles qui ont court pour la détection de cancers dans les IRM.

Quantib (2012, Pays-Bas, 5M€) propose Quantib Neurodegenerative, une solution à base de deep learning qui quantifie et segmente les IRM du cerveau.

Des chercheurs d'IBM à Yorktown et de la **fondation CHDI** à Princeton utilisent du deep learning pour faire de l'analyse tridimensionnelle de scans IRM du cerveau afin de détecter l'émergence de la maladie de Huntington, une maladie neurodégénérative malheureusement incurable[1271].

Max-Q.ai (2013, Israël, $2M) étudie les effets des accidents vasculaires cérébraux par l'analyse d'IRM en s'appuyant sur les briques de reconnaissance d'image d'IBM Watson, et semble-t-il, une une autre brique de détourage issue de **Mirada**. **Viz.AI** (2016, USA, $30,6M) a de son côté obtenu l'agrément FDA en février 2018 pour sa solution Contact qui réalise le diagnostic de l'artériosclérose cérébrale (vidéo).

[1269] Voir AI Detects and Measures Small Vessel Disease in Brain CT Scans, mai 2018.

[1270] Voir New AI Tool, HeadXNet, to Help Detect Brain Aneurysms par Siavash Parkhideh, juin 2019 qui fait reference à JAMA Network Open: Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model, juin 2019.

[1271] Voir Resting-state connectivity stratifies premanifest Huntington's disease by longitudinal cognitive decline rate par Pablo Poloseckiet al, 2020 (15 pages) et IBM uses AI to predict progress of Huntington's disease symptoms par Jon Fingas, janvier 2020.

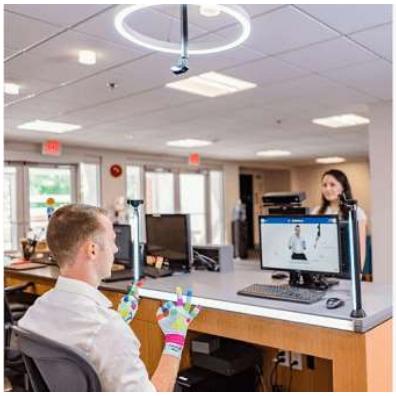
Selon des chercheurs de l'**Hôpital pour enfants de Boston**, la combinaison de la dilatation de la pupille et du battement cardiaque permettrait de diagnostiquer l'autisme plus tôt lors de la croissance des enfants ainsi qu'une autre pathologie rare d'origine génétique, le syndrome de Rett[1272]. Le tout exploite du machine learning qui fait des corrélations entre ces signaux et l'émergence de ce type de pathologie dès le plus jeune âge (6 à 18 mois). Les premiers modèles étaient bâtis avec des souris et ont ensuite été étendus à 35 jeunes filles atteintes du syndrome de Rett. Le taux de réussite de la détection était de 80%. Ils pourraient être améliorés en ajoutant des examens d'EEG aux modèles. Ici, nous sommes dans le phénotype.

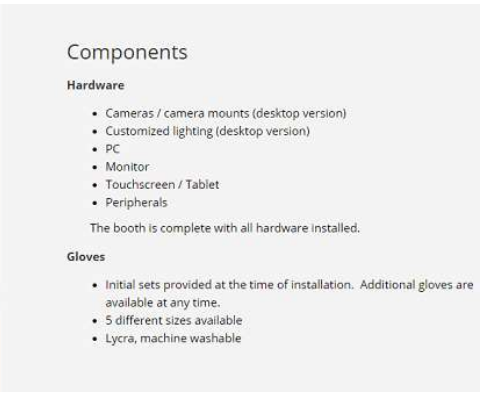
Passons au génotype…

En 2019, une équipe de chercheurs de l'**Université de Princeton** identifiait avec du deep learning des mutations des séquences non codantes de l'ADN pouvant modifier l'expression des gènes entraînant l'autisme. Cela fait partie du champ de l'épigénétique. L'analyse de l'ADN portait sur 1790 familles dont un seul des membres était atteint d'un syndrome d'autisme. Elle couvrait 120 000 mutations différentes pour identifier celles qui avaient une corrélation avec l'autisme. Cela complète les analyses génomiques portant sur les mutations des séquences codantes de l'ADN, celles qui sont dans les plus de 20 000 gènes connus qui servent à fabriquer les protéines dans nos cellules. Dans l'échantillon de patients analysés, seuls 30% des personnes atteintes d'autisme avaient des causes génétiques connues. L'épigénétique permet ainsi d'élargir ce champ de diagnostic et d'en éviter des erreurs[1273].

L'IA et l'analyse visuelle permettent de créer des solutions de lecture du langage des signes des sourds ou malentendants.

C'est ce que propose **SignAll** (2016, Hongrie, $3,6M) qui analyse visuellement la langue des signes américains pour le transformer en texte en clair. Aux USA, cette langue s'appuie sur l'usage d'une seule main[1274]. Le système analyse aussi la position du corps ainsi que les expressions faciales.





Il serait capable de reconnaître des phrases entières et pas seulement des signes isolés. C'est l'œuvre de réseaux à mémoire dérivés de LSTM. L'utilisateur doit tout de même enfiler des gants colorés, les couleurs servant de marqueurs pour la vision artificielle s'appuyant sur trois caméras. Le système a été entraîné avec 250 000 images labellisées. Sachant que le système doit être entraîné avec des données spatio-temporelles, pas juste des images statiques.

La startup est partenaire depuis 2017 de l'Université Gallaudet située à Washington D.C. Le marché cible ? Les 70 millions de sourds et malentendants dans le monde et surtout leur entourage familial et professionnel. L'inconvénient ? Cela ne fonctionne pas sur smartphone ni en mobilité et seulement pour la langue des signes utilisée aux USA. Depuis, un développeur a ajouté une fonction très rudimentaire de reconnaissance des signes à **Amazon Alexa** (vidéo) qui fonctionne signe par signe, équivalent du mot à mot, et via un ordinateur portable.

[1272] Voir Pupil dilation and heart rate, analyzed by AI, may help spot autism early, juillet 2019.

[1273] Voir Artificial intelligence detects a new class of mutations behind autism, mai 2019.

[1274] Voir Interpreting Sign Language Using Artificial Intelligence, août 2019. En France, le langage des signes utilize les deux mains.

Des chercheurs américains développaient en 2019 une IA qui améliore le diagnostic de la maladie d'Alzheimer[1275]. Elle sert à détecter les plaques d'amyloïde et les angiopathies cérébrales dans les biopsies du cerveau. Le système a été entraîné, comme il se doit, avec des images de biopsies dûment labellisées. En gros, c'est un outil d'anatomopathologie utile pour la recherche dans le domaine.

Dans la lignée de ce veut faire la startup israélienne **Beyond Verbal**, une équipe de chercheurs de la NYU School of Médecine (New York) a pu identifier les syndromes d'un PTSD (post-traumatic syndrom disorder) de vétérans de l'armées US en analysant leur voix. La détection est bonne à 89% ce qui est moyen et doit compléter d'autres méthodes de diagnostic. La méthode de machine learning utilisée s'appuie sur des arbres aléatoires (random forest) de classification. Le système de reconnaissance de la parole utilisé venait du SRI[1276].

Enfin, des chercheurs de l'**Université de Hong Kong** ont de leur côté automatisé des tests d'identification de différentes formes de démence. Une équipe associant Apple et l'**Université de Tübingen** en Allemagne (près de Stuttgart) a trouvé le moyen de détecter de troubles cognitifs de personnes âgées en se contentant d'analyser l'ordre dans lequel elles enchaînent l'usage d'applications mobiles. En effet, l'ordre d'utilisation des applications téléphone, messages, messagerie, navigateur et et calendrier se retrouve souvent chez les les personnes présentant des troubles cognitifs[1277]. L'échantillon de personnes âgées utilisé pour l'entraînement du système restait modeste avec 113 participants.

Le machine learning est employé pour analyser les résultats de tests psychotechniques comme les tests de Montréal (mémoire à court terme, richesse du vocabulaire, capacuté d'attention, nommer des animaux, dessiner le plan d'une horloge) et évaluer notamment les variations de résultats dans le temps[1278].

Diabète

L'IA intervient dans la gestion du diabète, nous le verrons plus loin, mais assez peu dans sa détection et sa prévention en tout cas pour le diabète type 1 insulinodépendant.

IBM et une association de diabétiques US annonçait en 2019 avoir créé une IA capable de prédire l'émergence d'un diabète type 1 à partir de la présence d'anticorps spécifiques. L'étude s'appuyait sur les données de patients récoltés aux USA, en Suède et en Finlande[1279]. Mais il faut là encore prendre cela avec des pincettes. Ces tests d'anticorps ne font pas partie des bilans sanguins standards et s'ils l'étaient, les chercheurs indiquent que l'IA générerait de nombreux faux positifs.

La mesure de glycémie en continu est maintenant courante comme avec le patch FreeStyle d'**Abbott** ou le **Dexcom 6**. Leurs résultats peuvent être exploités pour aider les patients à bien équilibrer leur diabète. Leurs dernières moutures sont reliées en temps réel et signaler des épisodes d'hypoglycémie ou d'hyperglycémie nécessitant un ajustement soit par la prise de glucose pour les premiers soit d'insuline pour les seconds. Mais sans IA. L'IA peut intervenir pour aider à évaluer ces épisodes de manière indirecte et via un ECG (électrocardiogramme)[1280].

Au CES 2016, IBM présentait avec l'équipementier médical **Medtronic** une autre solution utilisant Watson pour prédire la survenue d'hypoglycémies des diabétiques de type 1.

[1275] Voir Artificial Intelligence to Speed up Alzheimer's Disease Research, mai 2019.

[1276] Voir Artificial intelligence can diagnose PTSD by analyzing voices, avril 2019.

[1277] Voir Apple développe une IA capable de prédire les troubles cognitifs par Valentin Cimino, novembre 2019.

[1278] Voir Spoken Language Technology Takes on Dementia, septembre 2019.

[1279] Voir AI could be the key to catching Type 1 diabetes much earlier, juin 2019.

[1280] Voir AI Powers Personalized Medicine Approach to Detecting Hypoglycemia par Medgadget Editors, janvier 2020.

L'hypoglycémie est générée par une boucle de rétroaction qui associe l'activité physique, prise d'insuline et alimentation. Il faut donc mesurer les trois ce qui n'est pas trop compliqué pour les deux première mais moins évidente pour la dernière, même avec les capteurs infrarouge de type Scio (qui ne fonctionnent pas si bien…).

Cependant, l'application est probablement pertinente pour les diabétiques qui pratiquent un sport intensif et pour lesquels les risques d'hypoglycémie sont importants et répétés. Sachant que les algorithmes doivent tenir compte des effets retard de l'alimentation comme de la prise d'insuline.

L'IA peut aussi intervenir dans la mesure de la glycémie. C'est le cas dans le prototype de « radar » la captant codéveloppé par l'**Université de Waterloo** au Canada, Google avec son kit radar Soli et l'Allemand Infineon. Le radar utilise la bande de 60 GHz, dans la tranche haute des ondes millimétriques (la 5G va utiliser le 26 GHz). L'outil sert surtout à la détection de mouvements de proximité comme avec un LeapMotion qui fonctionne dans l'infrarouge. Avec un peu de traitement du signal et d'IA, les chercheurs ont réussi à mesurer le niveau de glycémie de manière totalement non invasive. Bon, pour l'instant, cela fonctionne en éprouvette mais pas sur de vrais gens [1281]!

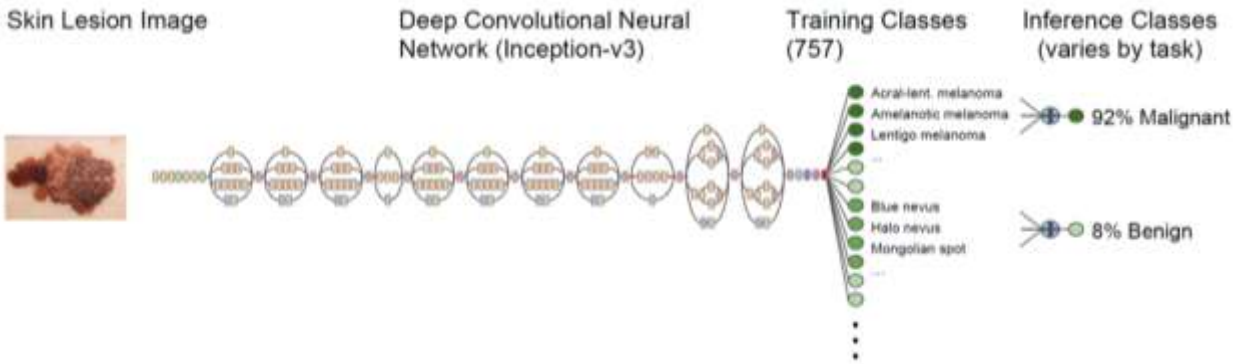
Reins

Nous avons déjà cité l'implication de **DeepMind** dans la détection de maladies rénales avec le NHS au Royaume-Uni. La filiale d'Alphabet récidivait en 2019 en annonçant pouvoir détecter rapidement les risques d'insuffisance rénale aiguë, requérant une dialyse[1282]. Une évaluation de cette IA a été réalisée avec l'administration des vétérans de l'armée US (VA) sur 700 000 dossiers de patients. Le système utilise une méthode d'anonymisation des données médicales des patients. Le taux de détection correct est de 90%. Mais ce n'est visiblement pas encore déployé.

Dermatologie

L'imagerie médicale peut également jouer un rôle en dermatologie pour rendre cette discipline plus accessible, et pourquoi pas pour certains médecins généralistes.

Andre Esteva et Sebastian Thrun de l'**Université de Stanford** décrivent très bien les enjeux et les méthodes utilisées dans leur méthode[1283]. La classification des lésions utilise une méthode assez classique avec un R-CNN pour les isoler puis un réseau convolutif, ici inception-v3[1284], pour détecter les formes et un classifieur permettant d'isoler 757 variantes de cas dont 92% sont des lésions malignes (*ci-dessous*).



[1281] Voir Artificial Intelligence and Radar Technologies to Measure Blood Glucose, juin 2018.

[1282] Voir DeepMind AI Will Be Able To Predict Acute Kidney Injuries 48 Hours Before It Happens par Tyler Lee dans Ubergizmo, janvier 2019 et The VA Wants to Use DeepMind's AI to Prevent Kidney Disease par Tom Simonite dans Wired, janvier 2019.

[1283] Dans la présentation Dermatologist-level classification of skin cancer with deep neural networks, 2017 (48 slides).

[1284] Voir Rethinking the Inception Architecture for Computer Vision, d'une équipe de Google, 2015 (10 pages). Ce modèle de réseau convolutionnel présente l'avantage d'être peu couteux en temps machine pour son entraînement.

Dans le même créneau, **SkinVision** (2012, Pays Bas, $12,2M) a développé une application mobile qui analyse les lésions sur la peau en trente secondes avec une sensibilité de 95% pour détecter les tumeurs malignes émergentes. Le tout avec la caméra du smartphone.

Une expérience concluante a été publiée en 2018 par une équipe de chercheurs allemands (Heidelberg, Göttingen, Konstanz, Passau), américaine (New York) et française (Centre de Recherche en Cancérologie de Lyon)[1285]. Leur IA à base de deep learning génère un meilleur diagnostic de cancers de la peau que 58 dermatologues de référence. Le système avait été entraîné avec 100 000 images labellisées. Le taux de bonne reconnaissance était de 95% pour l'IA pour une moyenne de 89% pour les dermatologues.

Imagerie généraliste

Les solutions d'imagerie utilisent à peu près toutes les mêmes techniques et solutions logicielles mais sont paramétrées de manière différente selon les pathologies recherchées et avec des jeux d'entraînement spécifiques.

Elles vont bien au-delà des techniques simples d'augmentation de contraste qui existaient auparavant. Elles détectent des formes particulières, des densités spécifiques et réalisent aussi des mesures précises et quantitatives.

Elles peuvent aussi comparer avec précision des images dans le temps pour quantifier l'évolution d'une pathologie ou de son traitement. Elles utilisent généralement différentes formes de réseaux convolutifs, avec au moins un premier qui détecte des formes dans l'image et un autre qui les labellise une par une après détourage.

Pour être entraînées, ces solutions doivent évidemment exploiter des bases d'images déjà labellisées par des spécialistes et issues de réseaux de laboratoires, cliniques et hôpitaux. Ces données ne sont pas forcément ouvertes et les startups doivent monter des partenariats ad-hoc pour les récupérer. Très souvent, les startups françaises se sourcent aux USA pour récupérer ces bases de données[1286].

Les rapports d'analyse de ces systèmes sont aussi rédigés avec des textes en clair, ce qui relève du traitement du langage[1287] avec des réseaux à mémoire (LSTM) et même des réseaux génératifs (GAN).

Une nouvelle discipline née en 2012 complète l'analyse d'images classique avec labellisation des tumeurs ou pathologies découvertes par des réseaux convolutifs, les **radiomics**[1288].

Il s'agit d'une méthode différente d'analyse des images qui génère des données quantitatives sur des « features » de ces images.

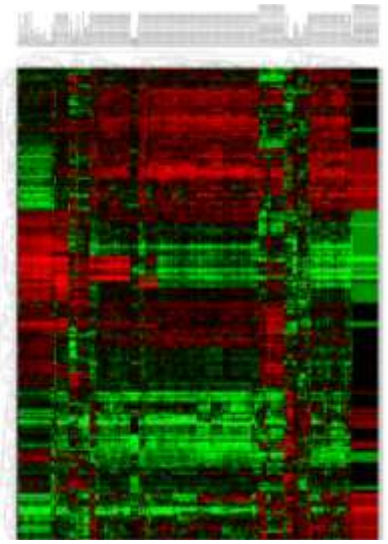
[1285] Dans Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, 2018 et relatée dans Artificial intelligence for melanoma diagnosis: how can we deliver on the promise?, août 2018.

[1286] A noter l'existence de **TeraRecon** (1999, USA), un fournisseur de bases d'imagerie médicale qui permet d'entraîner des IA. Ils disposent de centaines de millions d'examens y compris de l'imagerie 3D. Leur filiale WIA Corporation issue de l'acquisition de McCoy Medical Technologies en 2017 propose l'accès à des algorithmes d'IA via une plateforme et un jeu d'APIs. C'est en gros un intermédiateur d'algorithmes de traitement d'images.

[1287] Une méthode est décrite dans On the Automatic Generation of Medical Imaging Reports 2018 (10 pages).

[1288] Pour en savoir plus sur ce vaste sujet, voir Radiomics the process and the challenges, 2012 (15 pages), Radiomics for outcome modeling: state-of-the-art and challenges de Mathieu Hatt (INSERM), 2018 (60 slides), Radiomics Images are More than Meets the Eye, de Ralph Leijenaar, 2017 (197 pages), Machine Learning Applications for Radiomics Towards Robut Non-Invasive Predictors in Clinical Oncology de Chintan Parmar, 2017 (187 pages) et Radiomics applied to lung cancer: a review, 2018 (12 pages).

En gros, il s'agit de réaliser de l'analyse de données macro sur les images. Elle permet d'identifier des caractéristiques pathologiques qui ne sont pas détectées par l'analyse visuelle classique d'images et par réseaux convolutifs. Elle peut-être aussi croisée avec des données autres comme celles du génome ou des données phénotypiques pour identifier des corrélations. Cela génère des visualisations étonnantes comme *ci-contre*, avec dans les lignes, l'apparition de ces « features » dans les images en en colonne, des patients ayant ici un cancer du poumon.



De nombreuses startups se positionnent comme des généralistes couvrant plusieurs pathologies. Certains prévoient la fin des radiologues[1289]. Lorsqu'on lit entre les lignes, on se rend compte que les radiologues seront amenés à s'appuyer sur des IA à base de reconnaissance d'image pour les aider à interpréter les images.

On aura cependant toujours besoin de radiologues, ne serait-ce que pour labelliser les images qui entraînent ces systèmes de deep learning ! L'IA les aidera à être plus efficaces[1290]. Cela peut sembler bizarre mais comme les algorithmes de détection sont assez voisins d'une pathologie à l'autre, pourquoi pas créer des solutions multi-pathologies. Mais les modes de commercialisation ne sont pas les mêmes d'une spécialité à l'autre.

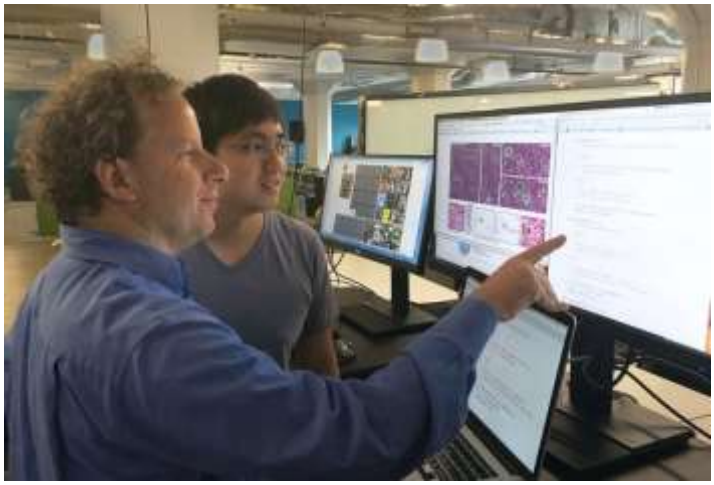
- **Butterfly Network** (2011, USA, $100M) créé un échographe dont toute l'électronique tient sur un seul composant, et dont les images sont analysées par « computer vision ». Il semble qu'il utilise plusieurs émetteurs ultra-sons, un peu comme le système d'OpnWatr qui fonctionne dans l'infrarouge pour faire de l'imagerie cérébrale. Son fondateur Jonathan Rothberg a de l'expérience, ayant créé et revendu deux sociétés de séquençage de l'ADN, 454 à Roche et Ion Torrent à Thermo Fisher.
- **VoxelCloud** (2015, USA, $28,5M) couvre le cancer du poumon, la rétinopathie diabétique, les maladies coronariennes et du foie. Bref, un peu de tout.
- **Incepto Medical** (2018, France, $5,6M) est un généraliste de l'analyse de radios, avec des solutions pour les scans du cerveau en urgence, les scans du poumon, la détection de fractures osseuses, celle d'occlusions du système digestif, des anévrismes de l'aorte ainsi que les mammographies. Ces outils d'analyse s'intègrent dans les workflows existants de traitement des radios.
- **Lunit** (2013, Corée du Sud, $20.5M) propose une solution logicielle de deep learning générique d'interprétation d'imagerie médicale, notamment de radios et qui semble commercialisée en OEM et focalisée sur les poumons.
- **Infervision** (2015, Chine, $70M) est un généraliste de l'analyse d'images médicales à base de deep learning. Ils ont déjà déployé leur solution dans 280 services d'oncologie dans le monde. Leur système améliorerait l'exactitude des analyses de 20% par rapport aux radiologues pour certaines pathologies[1291]. Le marché chinois a ceci de particulier que les déploiements massifs y sont plus rapides !

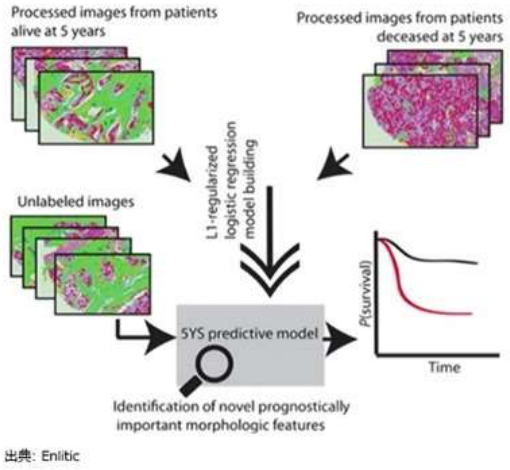
[1289] Voir Why Vinod Khosla thinks radiologists still practicing in 10 years will be 'causing deaths' par Darrell Etherington, juin 2019. Et un contre-point dans The Future of Medical Imaging and Machine Learning, avril 2019.

[1290] Voir Voici comment une ingénieure en IA imagine le métier de radiologue dans le futur par Chisato Goya, novembre 2019.

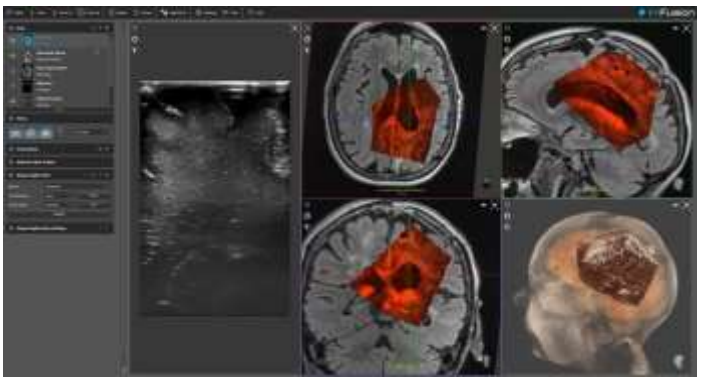
[1291] Voir China's Infervision is helping 280 hospitals worldwide detect cancers from images par Rita Liao, novembre 2018.

- **Enlitic** (2014, USA, $15M) propose de l'aide au diagnostic en s'appuyant sur les résultats de divers systèmes d'imagerie médical (IRM, scanner, radios) et sur du deep learning (*ci-dessous à gauche* avec son fondateur Jeremy Howard). Il détecte des pathologies émergentes le plus tôt possible, notamment les cancers du poumon. Il aide aussi à identifier plusieurs pathologies simultanément[1292].





- **ImFusion** (1998, Allemagne) fait de l'imagerie composite qui assemble des vues d'imagerie ultrasons et de scanners classiques (rayons X et IRM) pour créer des vues 3D qui sont ensuite exploitées par des réseaux de neurones et avec des serveurs Nvidia[1293]. Le tout permet d'exploiter de l'échographie réalisée pendant des opérations chirurgicales, superposée à des IRM/scans réalisés avant l'opération.



- **IBM** qui développe une panoplie de solutions d'interprétation d'imagerie médicale avec Watson Health et dont les solutions logicielles sont aussi exploitées par bon nombre de startups du secteur[1294].
- **SigTuple** (2015, Inde, $25M) fait de l'analyse d'échantillons sanguins à base d'imagerie avec sa solution Shonit (vidéo) qui fait de comptage de cellules. Cela peut notamment servir à diagnostiquer l'anémie et la malaria.

Ces IA d'analyse des radios vont continuer à s'améliorer grâce aux progrès des systèmes de radiologie eux-mêmes. Leur résolution va probablement s'améliorer grâce la capacité de gérer des systèmes avec une plus grande mémoire vive[1295].

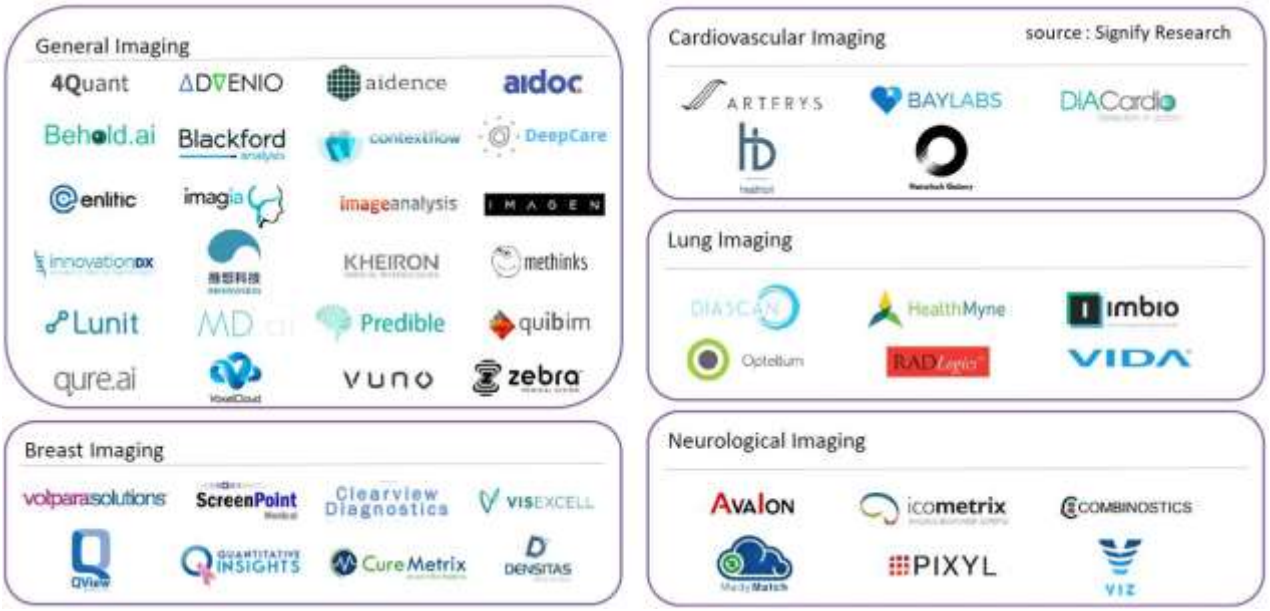
[1292] Voir la **vidéo** de son CEO, Jeremy Howard à TEDx Bruxelles en décembre 2014. Il y aborde un point clé : il n'y a pas assez de médecins dans le monde. L'automatisation des diagnostics est donc un impératif incontournable.

[1293] Voir AI Innovators: Healthcare Startup Repurposes Ultrasound Imaging par Mona Flores, 2018 et Munich Startup Uses AI to Take Medical Imaging to Another Dimension, par Daniel Saaristo, février 2019.

[1294] Voir cette intéressante analyse de la position d'IBM Watson dans l'imagerie médicale : http://www.nanalyze.com/2017/08/ibm-dominate-radiology-ai/ ainsi que IBM's Automated Radiologist Can Read Images and Medical Records de Tom Simonite, février E2016 dans la MIT Technology Review.

[1295] Voir Addressing the Memory Bottleneck in AI Model-Training for Healthcare, Dell-EMC (15 pages) qui porte sur les moyens d'entraîner des modèles de deep learning avec un grand nombre de paramètres nécessitant des centaines de Go de mémoire vive dans les serveurs.

L'imagerie médicale n'est pas la seule source de diagnostics médicaux. Il faut ajouter l'analyse d'EEG (électro-encéphalogrammes), les tests biologiques (sang, urine) ainsi que les tests d'ADN (génotypie et séquençage complet).



De nombreuses startups ambitionnent d'exploiter tout ou partie de ces données pour améliorer les diagnostics, surtout dans le cadre de médecine préventive et pas seulement curative.

HealthReveal (2015, USA, $11,3M) propose une solution en cloud de prévention de l'apparition de maladies chroniques liées au style de vie, basée sur l'utilisation de capteurs biométriques divers. La solution en cloud exploite ces données en plus des données de parcours de santé. es clients sont plutôt les tiers payants dans la santé, surtout sur le marché des USA.

Forward (2016, USA) est une étonnante startup qui veut réinventer le cabinet médical. Son premier site à San Francisco est équipé de tous les capteurs [1296], outils d'analyses de laboratoires, ADN compris, et systèmes d'imagerie médicale pour faire un bilan de complet à 360°[1297]. Ce n'est pas une clinique pour autant. L'offre qui est proposée sur abonnement se veut être une solution de médecine préventive.

[1296] Ils ont même développé leur propre scanner corporel qui capte à distance le pouls et la température corporelle. Mais cela ne remplace pas une IRM ou une tomodensitométrie.

[1297] Loïc Le Meur a filmé avec son smartphone une visite assez complète de Forward à San Francisco et c'est très instructif : https://www.facebook.com/loic/videos/381807855521818/.

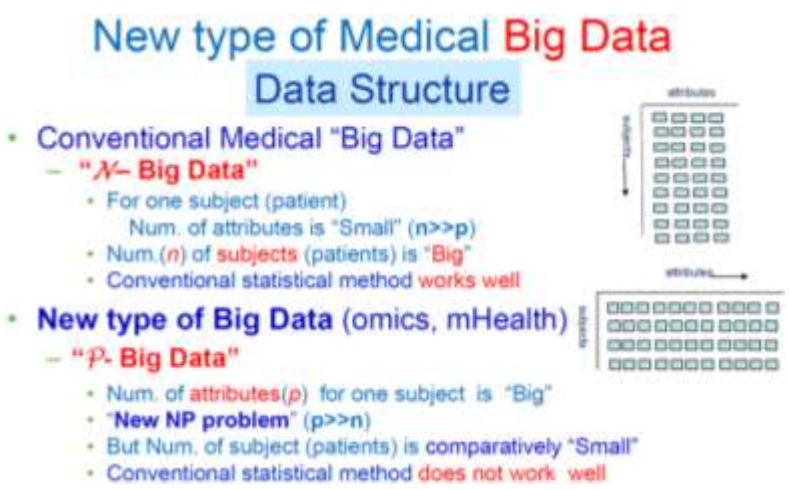
Enfin, il faudrait intégrer dans ce panorama l'émergence de la notion de **jumeau numérique** qui permet de simuler la réaction du corps humain à divers traitements[1298].

Génomique

La génomique est un domaine plein de promesses pour ce qui est de la médecine autant prédictive que curative. Nombre de startups se sont lancées dans la collecte du génotype ou du génome des patients, la différence entre les deux n'étant pas encore bien comprise par le grand public[1299]. Les données collectées dans les deux cas sont de grande taille pour chaque patient.

Ce qui fait qu'en termes de « big data », la génomique et toutes les autres « omics » (proteomics, bacteromics, etc), les outils d'analyse font appel à des données différentes.

Dans le big data traditionnel, on a un nombre limité de paramètres pour un grand nombre d'enregistrements, comme les caractéristiques de clients dans une base de clients. Dans les « omics », on a un très grand nombre de paramètres, qui s'évaluent en centaines de milliers à milliards, et avec un nombre généralement plus limité d'enregistrements (patients). Les plus grosses bases de génomes en comprennent aujourd'hui quelques dizaines de milliers[1300].



L'IA et le machine learning sont là pour exploiter tout cela ! Il faut cependant se méfier des études, notamment celles qui croisent génotype et phénotype, qui y trouvent des corrélations sans les expliquer[1301].

Deep Genomics (2014, USA/Canada, $16,7M) a créé le DG Engine qui analyse les variations du génome – les mutations de l'ADN – et la manière dont elles affectent le fonctionnement des cellules et génèrent des pathologies.

Ce sont des "genome-wide association study" (GWAS) qui produisent des analyses de corrélations entre modifications des gènes et pathologies (le "phénotype"). Les analyses réalisées par Deep Genomics ont la particularité d'intégrer tout le cycle de vie des gènes et notamment leur épissage – qui correspond à l'extraction de la partie codante des gènes – jusqu'à leur translation, à savoir la conversion de l'ARN qui résulte de l'épissage en protéines dans les ribosomes. Ils proposent en open source leur base de données SPIDEX de mutations génomiques et de leurs effets sur leur épissage[1302].

[1298] Voir Votre jumeau numérique pourrait bien vous sauver la vie par Coralie Lemke, décembre 2019.

[1299] Le génotype proposé par exemple par la startup 23andme consiste à évaluer les différences entre le génome standard humain et votre génome. Environ 500 000 différences sont évaluables par les techniques courantes de génotypage, qui ont un prix public d'environ $100. Le séquençage du génome complet consiste à analyser en détail les trois milliards de bases de l'ensemble de nos chromosomes, dont seulement 1,5% comprend la partie codante des gènes servant à générer des protéines. En gros, le séquençage génère beaucoup plus de données que le génotypage. Mais on ne sait pas encore très bien l'exploiter pour faire de la médecine prédictive et curative.

[1300] Le schéma vient de la présentation Big data and artificial intelligence in medicine and drug discovery de Hiroshi Tanaka (39 slides).

[1301] Voir Healthcare Needs AI, AI Needs Causality par Alexander Lavin, août 2019.

[1302] Voir The human splicing code reveals new insights into the genetic determinants of disease, 2015, qui explique les fondements scientifiques de leur procédé.

L'ambition est de mener à de la médecine personnalisée mais on en est encore loin. La société a été cofondée par Brendan Frey, qui avait fait son PhD à Toronto avec Geoff Hinton, un chercheur canadien à l'origine du décollage du deep learning en 2006 et qui est maintenant chez Google.

Pathway Genomics (2008, USA, $43M) propose divers tests génétiques et biopsies ciblés par risques pathologiques permettant d'identifier des facteurs de risques divers et variés en cardiologie (Cardia DNA Insight), dermatologie (SkinFit), BRCATrue (cancer du sein), ColoTrue (cancer du colon) et obésité (Healthy Weight DNA Insight). Le test Mental Health DNA Insight permet d'évaluer l'impact des traitements en psychothérapies et le Pain Medication DNA Insight évalue l'efficacité probable des analgésiques. La société utilise IBM Watson. Ici, on a surtout affaire à un bon packaging par pathologie car les données exploitées par ces différents tests proviennent généralement des mêmes analyses, comme la génotypie réalisée par 23andme qui analyse plus de 500 000 variations dans les gènes (génotypie à base de SNP, ou single nucleotid polymorphisms).

iCarbonX (2015, Chine, $200M[1303]) veut faire des données médicales, génomiques comprises, une plateforme pour des applications en propre et tierce-partie.

La société collecte les données de santé de patients à 360° : génomique, phénotype, bactérome, protéome, données comportementales et psychologiques (on est en Chine…). Le tout exploite de l'IA dont la nature n'est aucunement précisée, tout comme n'est pas détaillée l'origine des données exploitée. Comment les patients fournissent-ils leurs données à iCarbonX ? Ils ont déjà une demi-douzaine de partenaires applicatifs dans leur Digital Life Alliance qui exploitent leurs données.

Avec **SomaLogic** (1999, USA, $395M, analyse biologique de protéines), **HealthTell** (2010, USA, $40M, évaluation de la réaction immunologique), **PatientsLikeMe** (2004, USA, $127M, réseau social de santé), **AOBiome** (2013, USA, $38,7M, thérapies probiotiques), **GALT** (General Automation Lab Technologies, 2014, USA, $10,5M, analyses du microbiome) et **Robustnique** (2010, Chine, cosmétique à base d'enzymes). Le dernier est **Imagu Vision Technologie** (2005, Israël) et a été acquis en 2016. Ils apportent l'analyse d'imagerie médicale à iCarbonX, permettant de compléter le profil santé des patients. Ce qui leur a permis, ipso-facto, d'établir un laboratoire de R&D en Israël.

Sophia Genetics (2011, Suisse/USA, $250M) piste les mutations de l'ADN des patients de plus de 400 hôpitaux pour améliorer la compréhension de cancers et de certaines maladies rares et la création de traitements personnalisés. Ils font appel au machine learning dans leurs processus.

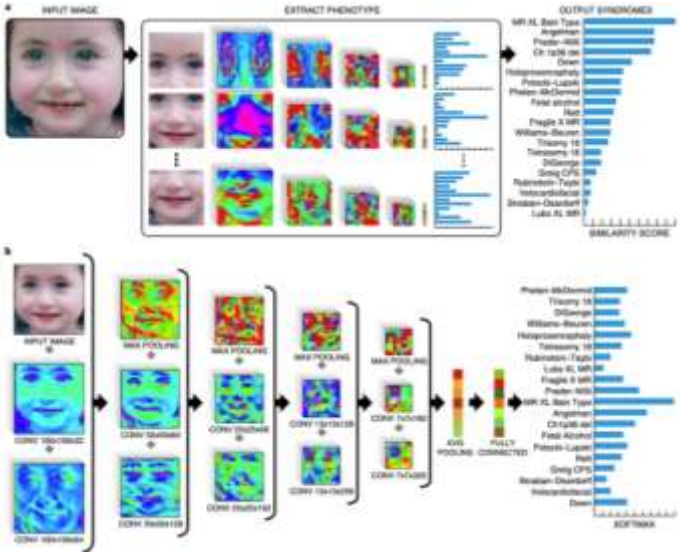
Ginger.io (2011, USA, $28,2M) a créé un outil de diagnostic et de prescription de traitement pour diverses pathologies neuropsychologiques. Il exploite des applications mobiles pour le diagnostic et du machine learning. La solution permet un auto-traitement de certaines pathologies par les patients.

Lumiata (2013, USA, $31M) est dans la même lignée un système d'analyse de situation de patient permettant d'accélérer les diagnostics, notamment en milieu hospitalier.

[1303] Ils annoncent avoir levé $600M, mais $400M correspondent à des investissements chez leurs partenaires. Voir Chinese AI company plans to mine health data faster than rivals, dans Nature, janvier 2017.

MedWhat (2010, USA, $3,2M) propose une solution générique d'aide au diagnostic qui s'appuie sur la panoplie totale de l'IA (deep learning, machine learning, NLP) et sous la forme d'une application mobile faisant tourner un agent conversationnel à qui on communique ses symptômes, qui pose des questions de qualification et oriente ensuite le patient (vidéo de démo). Elle stocke aussi le dossier médical du patient.

Citons une manière originale de croisement du génotype et du phénotype qui consiste à détecter des maladies génétiques rares par l'analyse du visage. C'est un projet de recherche dénommé DeepGestalt lancé par **FDNA** (2011, USA/Israel). L'entraînement était basé sur 17 000 images couvrant 200 syndromes [1304]. Il était dans un premier temps entraîné avec des visages d'enfants et plutôt caucasiens, entraînant un biais de faux négatifs pour les personnes de couleur, un défaut que les créateurs ont reconnu et annoncé vouloir corriger en élargissant leur base d'entraînement.



Dans les applications santé d'IBM Watson, on peut citer l'application de **GenieMD** (2010, USA, $100K) qui permet aux patients, aux USA, de faire un premier niveau d'autodiagnostic de problèmes de santé courants et d'être ensuite mis en relation avec des praticiens.

Il permet aussi de suivre l'observance de la prise de médicaments. La solution exploite les informations fournies par les patients en langage naturel. C'est une application générique qui pourrait être mise en oeuvre dans les stations de télémédecine pour les déserts médicaux.

Télémédecine

La télémédecine s'appuie sur différents outils à base d'IA. Il s'agit principalement d'intégration des outils que nous avons déjà vus, complétés par des outils de communication à distance.

Babylon Health (2013, UK, $85M) est une startup qui propose la combinaison de l'accès à un chatbot médical pour le diagnostic de premier niveau puis à un docteur en ligne pour poursuivre la discussion dans une visioconférence. Le diagnostic de premier niveau serait meilleur que celui des généralistes débutants au Royaume Uni (81% vs 72%). Le chatbot s'appuie sur un moteur de NLP pour comprendre les questions du patient et dialoguer avec lui, un Knowledge Graph pour modéliser les connaissances, un moteur d'inférence (système expert) pour réaliser son diagnostic, sachant qu'il est combiné à des modèles probabilistes et du machine learning.

[1304] Voir Facial recognition and AI could be used to identify rare genetic disorders, par James Vincent, janvier 2019.

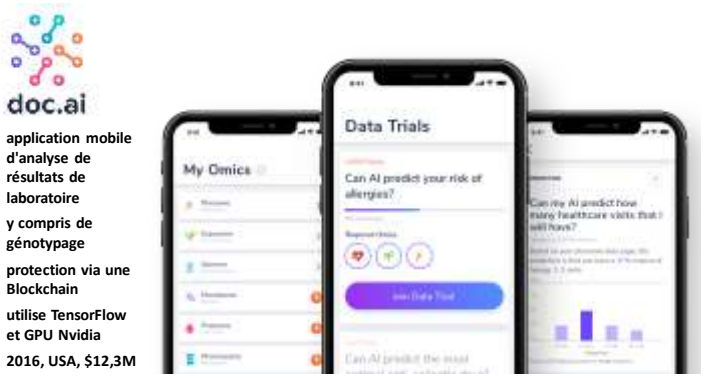
Une solution équivalente est développée par **Anamnèse** (2017, France, 500€). C'est un outil de téléexpertise médicale qui interroge les patients et interpréter ses symptômes afin de les orienter ensuite dans les systèmes de santé. Il utilise une combinaison de moteur de règles et de machine learning.

En Chine, le robot médecin **Xiaoyi** serait capable depuis fin 2017 d'égaler les médecins généralistes dans le pré-diagnostic, surtout en hôpital. Il aiderait les praticiens à remplir plus rapidement les dossiers des patients. Mais l'histoire ne dit pas à quelles informations des patients il a accès (source).

Et si le machine learning permettait d'améliorer la prise de vaccins dans un environnement où les anti-vaccins se multiplient, faisant réémerger les risques d'épidémies de maladies infectieuses comme la rubéole et la rougeole ? C'est un projet conduit par l'Université de Chicago qui s'est intéressé à un échantillon de 48 000 enfants d'origine croates entrant en école primaire dans différents pays dont la France, le Portugal et les USA entre 2011 et 2018. Ils ont utilisé un algorithme de régression logistique de type LASSO s'appuyant sur 25 paramètres et générant une fiabilité de 72%.

Est-ce généralisable au-delà des enfants croates ? Cela mérite réflexion sur les biais des données d'entraînement ! Ce genre d'algorithme risque de focaliser l'attention sur certaines communautés ethniques ou religieuses et générer un tollé associé[1305] !

Doc.ai (2016, USA, $41M) propose une application mobile et webapp comprenant un médecin virtuel dédié à l'analyse de résultats de laboratoires, y compris de génotypage, ce qui n'empêche pas, ensuite, d'aller voir un médecin et surtout un spécialiste (vidéo). Ils utilisent aussi une Blockchain pour gérer la traçabilité des examens biologiques et leur partage dans des communautés de patients pour des tests cliniques ou autres besoins.



La startup semble s'être repositionnée dans l'accompagnement de la recherche en exploitant les données captées par leur application.

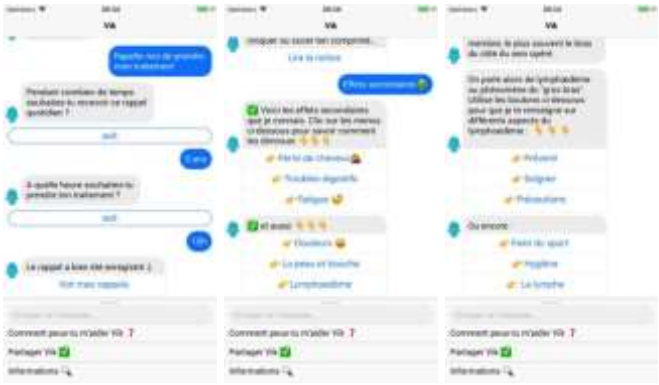
Medicus.ai (2015, Autriche, 6,6M€) analyse les résultats d'analyses biologiques de laboratoires pour générer des rapports compréhensibles par les patients. La société propose à la fois une application destinée aux patients avec divers tableaux de bord de reporting et une pour les entreprises, assurances et mutuelles santé. Cela fait penser à ce que faisait doc.ai à son lancement. Leur solution fait appel à du machine learning mais il n'est pas pour autant évident de détecter où précisément. En France, le dossier médical partagé pourrait exploiter cette fonctionnalité.

Karius (2014, USA, $254M) propose une solution de diagnostic de maladies infectieuses par l'analyse de biopsies liquides du sang, exploitant du séquençage de génome. L'exploitation des données générées par le séquençage s'appuie sur une base de 300 000 pathogènes et du machine learning.

[1305] Voir Machine Learning Predicts Kids at Risk of Not Getting Vaccinated par Jeremy Hsu, mai 2019.

Wefight (2017, France) a créé Vik Sein, un chatbot de suivi du cancer du sein pour les patientes (vidéo).

Il a été codéveloppé avec le réseau de proximité « Mon réseau cancer du sein ». Ce chatbot semble est assez riche fonctionnellement avec un suivi des traitements, des conseils et plusieurs types d'interactions avec les patientes.



Corti (2016, Danemark) a développé un outil d'analyse de la voix des patients ou de leur entourage qui appellent les services d'urgence pour détecter notamment les pathologies cardiaques et mieux gérer les urgences. C'est original et bien utile ! L'outil Orb alimente l'écran du dispatcher des urgences en l'alimentant avec les informations détectées dans la voix, y compris issues du speech-to-text. Orb utilise des réseaux convolutifs (vidéo).

Médecine prédictive généraliste

Prognos (2010, USA, $43,2M) est l'un des acteurs clés qui se positionnent dans le champ de la médecine prédictive. Ils ambitionnent de détecter le plus en amont possible une cinquantaine de pathologies en s'appuyant sur une gigantesque base de données de plus de 13 milliards de résultats d'examens biologiques de plus de 160 millions de patients. Le tout doit s'appuyer sur force machine learning avec analyses de corrélation entre paramètres biologiques et occurrences de pathologies. La startup annonce avoir créé plus de 1000 algorithmes propriétaires, et visiblement, sans données génomiques.

Medopad (2011, UK, $32M) adopte une approche voisine en collectant un maximum d'informations issues d'objets connectés comme les montres intégrant un capteur cardiaque, les trackers d'activité et autres lecteurs de glycémie ou de tension pour les analyser et détecter des états de santé nécessitant un traitement. Leur solution est commercialisée tout azimut, y compris aux assurances santé privées qui peuvent mettre en place des bonus/malus en fonction du comportement des assurés.

InsightRX (2015, USA, $2,8M) utilise le machine learning et les données patients pour optimiser le choix des thérapies et leur dosage. Les données comprennent cette fois-ci les informations de génomique. La solution serait déjà déployée dans des hôpitaux aux USA.

On apprenait aussi en mars 2019 qu'une IA pouvait prédire les **risques de décès prématurés**[1306]. Cela fait froid dans le dos ! Mais il y un peu tromperie dans le message simpliste délivré !

[1306] Voir Artificial intelligence can predict premature death, study finds, mars 2019 qui fait référence à Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches par Stephen Weng & Al, mars 2019.

L'étude associée sert surtout à montrer que cette prévision, qui s'appuie sur l'exploitation de très nombreux paramètres sur 500 000 habitants du Royaume-Uni dont 14 500 sont décédés, est meilleure avec un algorithme de deep learning (76%) qu'avec une méthode classique de machine learning à base de forêt aléatoire (« random forest », 64%) ou une méthode de régression classique de Cox (44%). Une méthode à base de deep learning présente l'intérêt de mieux combiner les probabilités croisées de facteurs de risques. Mais 76% d'efficacité reste moyen et rappelle que l'avis de praticiens reste utile. La médecine probabiliste n'a rien à voir avec la médecine personnalisée !

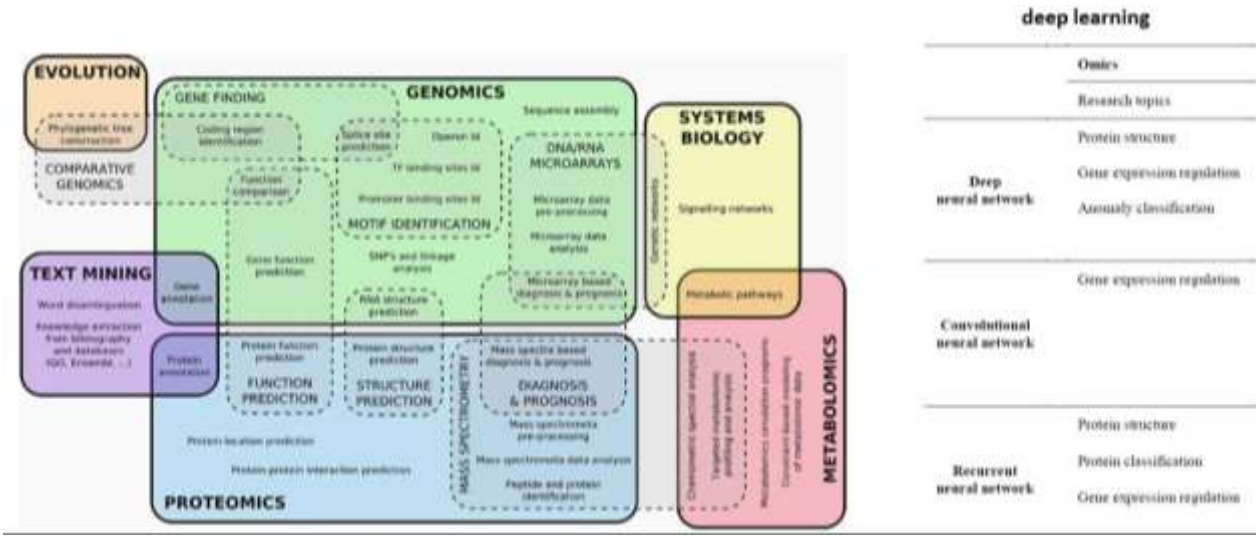
Enfin, on pourrait aussi **analyser le cri des bébés** pour détecter automatiquement les raisons de ces cris et notamment si les bébés ont faim ou s'il faut changer leurs couches[1307]. Le genre de solution typique du syndrome du « *comment faisait-on avant ?* ». C'est destiné à la fois aux parents et aux professionnels qui s'occupent des bébés (crèches, hôpitaux…).

Thérapies

Passons maintenant à l'exploitation d'IA pour guérir !

Les biotechs sont de grandes consommatrices de logiciels et d'IA. En cause, les volumes de données à gérer et analyser, à commencer par ceux qui viennent de toutes les techniques en « omique » : la génomique (analyse de l'ADN et de l'ARN) et la protéomique (analyse des protéines). La baisse du coût du séquençage de génomes de toutes les espèces vivantes a généré d'énormes quantités de données à exploiter.

L'IA peut aider à comprendre la structure des gênes et de leur expression, l'épissage des gènes (comment les différentes parties d'un gène s'assemblent), le repliement des protéines sur elles-mêmes après leur production dans les ribosomes des cellules ou la détermination des paramètres qui favorisent ou pas l'expression des gênes, notamment ceux qui déclenchent des cancers[1308].



Le schéma *ci-dessus* illustre la variété des usages de l'IA dans ces domaines (source). Les techniques employées tournent essentiellement autour du machine learning. Le deep learning commence aussi à faire son apparition dans certains cas d'usages listés *ci-dessus*[1309].

[1307] Voir Infant Cry Language Analysis and Recognition: An Experimental Approach par Lichuan Liu, mai 2019. Les travaux viennent de chercheurs chinois installés aux USA.

[1308] Voir Deep Learning in Pharmacogenomics From Gene Regulation to Patient Stratification 2017 (40 pages) qui fait un bon tour d'horizon de méthodes de deep learning utilisées dans la recherche en génomique.

[1309] Voir Deep Learning in Bioinformatics des Coréens Seonwoo Min, Byunghan Lee et Sungroh Yoon, 2016 (40 pages).

Le deep learning permet de réaliser des prévisions de comportements de molécules et de structures de protéines, des problèmes mathématiques très complexes à résoudre par des méthodes traditionnelles[1310]. Et nous n'en sommes qu'au début dans ce domaine, l'informatique quantique pouvant à plus long terme permettre d'aller encore plus avant dans ces simulations.

L'IA ne permet cependant pas pour autant de créer cette fameuse « médecine de précision » 100% adaptée à chaque individu et en fonction de son génome. C'est pour l'instant bien trop complexe, ne serait-ce que d'un point de vue biologique[1311].

Bien qu'utilisés essentiellement en imagerie médicale, les réseaux de neurones convolutifs sont aussi exploités en génomique. Mais ce sont les réseaux de neurones récurrents (RNN) qui sont plus couramment employés, car ils sont adaptés à l'analyse de données séquentielles comme pour le langage, or l'ADN est un langage, à base de quatre lettres (ATCG).

Nous ne traiterons pas ici du sujet dans ses détails mais allons plutôt l'illustrer par quelques startups ou chercheurs actifs dans le domaine.

Drug discovery

L'usage de l'IA dans la découverte de nouvelles molécules thérapeutiques est un champ d'exploration relativement nouveau et en pleine phase de maturation. Le machine learning peut aider à exploiter de gros volumes de données pour faire ressortir des molécules intéressantes, notamment par la méthode du criblage, qui associe des molécules connues à différentes cibles thérapeutiques. Au passage, cela permet de potentiellement raccourcir le cycle de mise au point de nouvelles thérapies[1312].

D'autres essayent de modéliser la structure tridimensionnelle des molécules organiques (peptides, polypeptides, protéines, enzymes) pour créer ou identifier des zones actives et même de simuler la dynamique moléculaire, c'est-à-dire l'évolution dans le temps de ces structures tridimensionnelles. Cela comprend aussi le vaste champ de la simulation du repliement des protéines. Il existe des méthodes s'appuyant sur du crowdsourcing comme via le site FoldIt et d'autres qui exploitent du deep learning. Enfin, des méthodes de simulation de la physique quantique régissant les interactions atomiques sont envisageables avec le calcul quantique.

Fin 2018, DeepMind décrivait **AlphaFold,** sa solution maison de repliement des protéines[1313]. Il publiait une nouvelle mouture encore plus avancée en novembre 2020, AlphaFold 2, améliorant d'environ 25% la précision de la modélisation 3D, le tout entraîné avec la structure de 170 000 protéines connues[1314]. Ce n'est cependant pas encore une solution parfaite et universelle. La modélisation 3D est approximative et elle fonctionne bien pour des protéines ressemblantes à celles de la base d'entraînement. Il est difficile de simuler avec des protéines qui déviraient significativement de celles de la base d'entraînement.

C'est en pareil cas que des simulations sur calculateurs quantiques auraient un intérêt, modulo le fait que pour ce faire, il faudrait disposer d'un très grand nombre de qubits, ce qui ne sera pas possible avant pas mal d'années voire décennies.

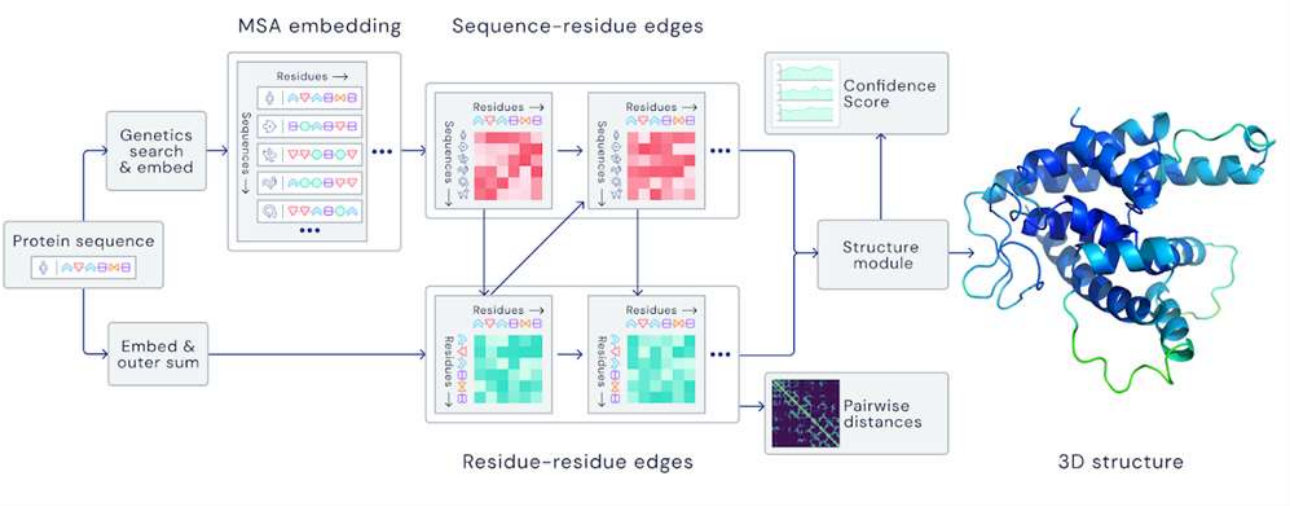
[1310] Voir Scientists develop machine-learning method to predict the behavior of molecules, octobre 2017.

[1311] Voir cet excellent papier qui remet les pendules à l'heure sur la médecine de précision : Why Does the Shift from "Personalized Medicine" to "Precision Health" and "Wellness Genomics" Matter? d'Eric Juengst et Michelle McGowan, septembre 2018.

[1312] Voir Pourquoi l'IA va révolutionner la pharmacologie moléculaire par Marc Rameaux, 2019, qui présente très bien la manière dont les techniques à base d'IA font gagner du temps dans la création puis la diffusion de nouvelles thérapies.

[1313] Voir Google's DeepMind aces protein folding par Robert F. Service, décembre 2018 et l'article de Nature correspondant : AI protein-folding algorithms solve structures faster than ever par Matthew Hutson, juillet 2019.

[1314] Voir AlphaFold: a solution to a 50-year-old grand challenge in biology, novembre 2020.



En 2018, des prouesses voisines à celles de la première version d'AlphaFold étaient annoncées par le biologiste Mohammed AlQuraishi de l'Université de **Harvard**[1315]. Des chercheurs allemands prédisent non pas la structure précise mais la fonction d'une protéine dans le corps humain avec le réseau de neurones **DeeProtein**[1316]. Celui-ci sert à identifier les parties de la protéine qui ont une fonction chimique précise.

Mais le deep learning n'est pas la seule solution. Elle peut aussi passer par de la simulation de l'interaction entre les composantes des protéines, comme cela a été réalisé début 2019 par des chercheurs au Texas et en Californie, s'appuyant sur un supercalculateur, et pour créer une structure moléculaire simple inspirée par celle de méduses[1317].

Il existe même des méthodes de création de nouvelles molécules thérapeutiques exploitant un réseau de neurones génératif comme **ReLeaSE**, créé par l'Université de Caroline du Nord[1318]. Enfin, à plus long terme, on pourra peut-être le faire via des algorithmes quantiques sur des ordinateurs quantiques disposant d'un très grand nombre de qubits intriqués.

Un bon nombre de chercheurs et de startups se positionnent maintenant sur ce créneau du « drug discovery »[1319].

Innoplexus (2011, Allemagne/Inde) propose iPlexus, un moteur de recherche d'informations médicales qui exploite 27 millions de publications, 365 000 rapports de tests cliniques et un million de thèses. Le tout s'appuie sur du machine learning pour générer des analyses quantitatives et du traitement du langage pour les recherches. La solution exploite aussi une Blockchain pour assurer la traçabilité des données intégrées dans la base. En plus de la pharmacie, la startup vise aussi les marchés financiers.

[1315] Voir AI's newest medical help is a new grasp of the protein-folding process par Stephen Shankland, 2019.

[1316] Voir What artificial intelligence can teach us about proteins, mai 2019.

[1317] Voir Supercomputers help supercharge protein assembly, mars 2019.

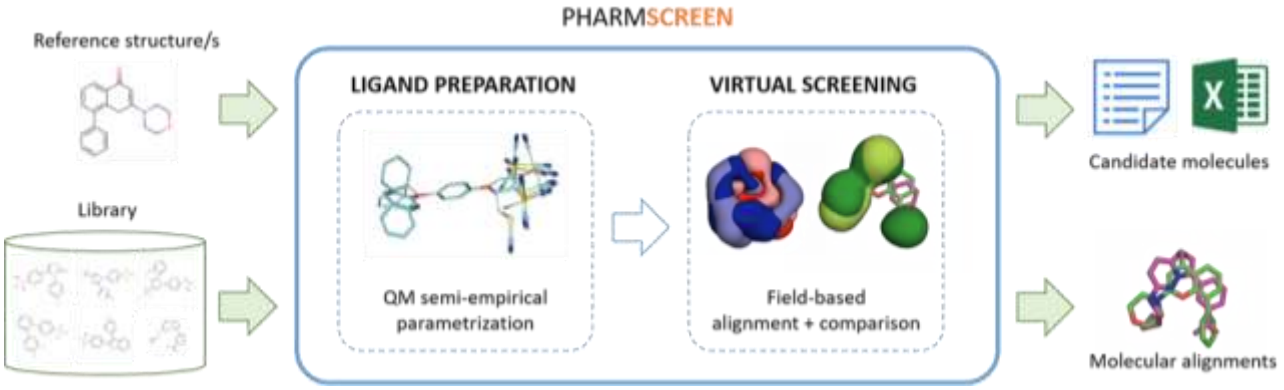
[1318] Voir Artificial intelligence system designs drugs from scratch, juillet 2018 et Deep reinforcement learning for de novo drug design, 2018 (27 pages).

[1319] Voir Using Artificial Intelligence for Drug Discovery, Nanalyze, avril 2019.

Iktos (2016, France) utilise du deep learning pour réaliser des simulations biologiques de l'effet de médicaments. L'idée consiste à screener des molécules existantes et à identifier in-silico leurs interactions avec des protéines connues selon un cahier des charges donné d'attaques de cibles à des fins thérapeutiques. Ils exploitent pour cela un réseau de neurones qui converti la structure des molécules connues dans un langage intermédiaire qui est ensuite rapproché des protéines cibles[1320].

C'est un mix de *stacked autoencoder* et de modèles génératifs. Leur solution permet de créer des molécules « in-silico ». La startup est partenaire avec le laboratoire pharmaceutique **Janssen**. Ils ne sont pas seuls sur ce marché qui comprend d'autres startups telles que Certara, ChemAxon, Mind the Byte, Optibrium et Tripos.

Pharmacelera (2015, Espagne, 290K€) a développé Pharmascreen, un moteur de recherche de molécules en exploitant des algorithmes de modélisation à base d'une IA, de nature non précisée.



Numerate (2007, USA, $19,4M) utilise l'IA pour aider à la découverte de nouveaux traitements à base de petites molécules pour traiter divers cancers, maladies neurodégénératives et cardiaques (ce qui couvre un bel éventail des besoins du marché). La startup travaille avec Merck. **Exscientia** (2012, UK, $107,9M) est positionné sur le même créneau.

Insilico Medicine (2014, Hong-Kong/USA, $51M) fait partie du grand nombre d'acteurs qui cherchent à trouver de nouvelles solutions curatives contre le cancer et les maladies du vieillissement à base de génomique et de big data. C'est en fait un prestataire de services qui créé de nombreuses solutions ad-hoc à base de deep learning. Il aide notamment d'autres entreprises à identifier de nouvelles thérapies, comme Pharmaceutical Artificial Intelligence. Leur logiciel en ligne aging.ai vous permet de déterminer votre âge à partir de vos résultats d'analyse sanguine. Mais vous pouvez aussi vous rappeler de votre date de naissance, ou au pire, consulter votre carte d'identité dans le pire des cas, ce qui sera probablement plus fiable. Ils ont aussi développé une application à base de deep learning qui permet d'identifier des traitements contre certaines fibroses[1321].

[1320] Une méthode qui semble voisine a été conçue par une équipe du MIT pilotée par Wengong Jin en et Regina Barzilay en 2018. Elle s'appuie sur une modélisation des molécules organiques sous forme de graphes entraînée avec 300 000 molécules différentes connues. Le projet a éré mené dans le cadre du consortium Machine Learning for Pharmaceutical Discovery and Synthesis piloté par le MIT et des entreprises de pharmacie, et financé en partie par la DARPA. Voir Junction Tree Variational Autoencoder for Molecular Graph Generation, 2018 (17 pages).

[1321] Voir This Startup Used AI To Design A Drug In 21 Days par Alex Knapp, 2019.

Envisagenics (2014, USA, $4,6M) réalise des analyses d'épissage de l'ARN pour détecter les cancers provoqués par la malformation de protéines et faciliter le développement de nouvelles molécules thérapeutiques qui corrigent ces problèmes d'épissage (vidéo). L'IA exploitée dans ce genre de solution n'est pas précisée.

Atomwise (2012, USA, $176,6M) utilise le machine learning pour découvrir de nouveaux médicaments et vérifier leur non toxicité. Le principe consiste à simuler l'interaction entre des milliers de médicaments connus et une pathologie telle qu'un virus, et d'identifier celles qui pourraient avoir un effet par simulation des interactions moléculaires. Un premier résultat aurait été obtenu en 2015 sur un virus d'Ebola. La simulation in-silico permet de choisir quelques médicaments qui sont ensuite testés in-vitro avec des cellules humaines.

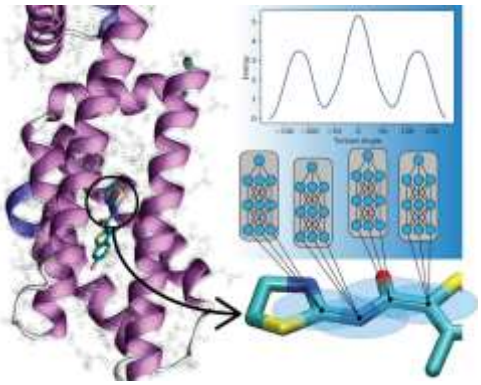
X37 (2018, USA,$14,5M) utilise aussi le machine learning pour la découverte de nouvelles thérapies ciblant notamment les maladies auto-immunes, certains cancers et un facteur anticoagulant. Ils exploitent AtomNet, la brique logicielle développée par **Atomwise**, pour découvrir des molécules simples permettant de cibler des protéines spécifiques.

Owkin (2016, France/USA, $18,1M) est sur le créneau du « drug repositioning » (ou retargeting, ou reciblage thérapeutique) qui permet d'étudier des essais cliniques et d'évaluer l'intérêt de certains médicaments sur d'autres pathologies que celles qui ont été testées. Ils ont aussi développé une solution de diagnostic du mésothéliome, un cancer s'attaquant à la plèvre qui entoure les poumons et qui est provoqué par l'exposition à l'amiante[1322]. Le tout s'appuie sur du machine learning. Ils sont partenaires de l'INSERM et de l'APHP, pour étudier les maladies du foie et du rein pour ces derniers.

HealX (2014, UK, $67,9M) fait aussi du reciblage thérapeutique à base d'IA, en ciblant les maladies rares. Ils utilisent surtout des techniques avancées de traitement du langage pour extraire des informations dans les publications scientifiques.

Pharnext (2007, France, $37M) fait aussi du criblage de molécules avec du machine learning pour identifier des possibilités de reciblage. La startup a été cofondée par Daniel Cohen, Philippe Pouletty et Serguei Nabirotchkin. Le premier est connu pour avoir créé la première cartographie complète du génome humain en 1993, avec son séquençage exhaustif autour de 2001 dans le cadre du Généthon[1323]. Ils sont en test clinique de phase III d'un traitement de la maladie de Charcot identifié par ce biais.

Des chercheurs de Los Alamos, de Caroline du Nord et de Floride utilisent de leur côté un réseau de neurones et de l'apprentissage par transfert pour décrire la cinématique des molécules lors de réactions exothermiques, d'isomérisations ou de modifications diverses exploitant les équations de la mécanique quantique, dans **ANI-1ccx**. Le réseau de neurones est entraîné avec la description de millions de molécules connues[1324].



[1322] De nombreux spots TV publicitaires aux US font référence à cette maladie alors qu'elle ne toucherait que 3000 nouveaux patients par an. Ces annonces dirigent les patients ou leur famille vers des avocats pour la réclamation de dommages et intérêts auprès d'industriels et employeurs en cause dans l'exposition à l'amiante. Ils créent un sentiment d'urgence lié, entre autres, aux statuts de limitation juridique sur les dommages et intérêts.

[1323] Voir How A.I. Is Finding New Cures in Old Drugs par Tiernan Ray, mars 2019.

[1324] Voir Researchers cast neural nets to simulate molecular motion, juillet 2019.

Standigm (2015, Corée du Sud, $3,7M) utilise le machine learning pour réduire les risques d'échecs lors de la phase I de tests cliniques de nouvelles thérapies hybrides associant plusieurs thérapies existantes[1325].

Dexstr.io (2014, France) est une startup toulousaine fournissant la solution Inquiro qui exploite les données médicales non structurées pour faciliter la recherche d'informations pour les sociétés de pharmacie. En gros, c'est de la recherche documentaire, un peu comme le font Sinequa et Antidot, mais avec un système adapté à la documentation scientifique dans la santé.

Autre phénomène, la combinaison du machine learning pour identifier des candidats thérapeutiques ou des variantes de candidats et de robots qui pilotent les tests in-vitro et simplifient les tâches répétitives. C'est ce que propose notamment **Recursion Pharmaceuticals** (2013, USA, $465M)[1326] ainsi que **LabGenius** (2012, UK, $28,7M).

Insitro (2018, USA, $246M) produit des thérapies géniques personnalisées combinant l'analyse du génôme des patients, des cellules souches, de l'édition du génome, et des tests in-vitro pilotés par du machine learning. Ils traitent notamment la stéato-hépatite non alcoolique (vidéo).

Berg (2006, USA) utilise de l'IA pour faire de la recherche thérapeutique en biotechs. Leur IA croise des données biologiques et phénotypales (environnement…) des patients. Ils exploitent diverses données : génomiques, protéomiques, lipidomiques, et métabolomiques, ainsi qu'au niveau du fonctionnement du cycle énergétique dans les mitochondries[1327]. Leur système de simulation « in silico » permettrait de simuler l'efficacité de thérapies, notamment contre les cancers et les maladies neurodégénératives. Ils ont créé une molécule qui restaure le processus de l'apoptose (la mort cellulaire programmée) de cellules cancéreuses (BPM 31510). Elle est en tests cliniques. Ils sont partenaires, entre autres, de Sanofi Pasteur en France pour développer un vaccin antigrippal fonctionnant d'une année sur l'autre ainsi qu'avec Astra Zeneca sur le traitement de la maladie de Parkinson.

Justement, le deep learning a été utilisé pour créer un vaccin contre la **grippe** par un chercheur australien. Les vaccins suggérés par le système ont été testés sur des cellules humaines après des tests animaux. Un vaccin retenu doit être testé aux USA pour la saison 2019/2020. Le procédé utilisé pour le concevoir n'a visiblement pas encore été publié en détail.

Il semble qu'il soit protégé pour être exploité par une startup biotech crée par les chercheurs australiens[1328]. Le machine learning peut aussi servir à découvrir de nouveaux antibiotiques[1329].

On peut aussi prédire certaines contre-indications de combinaisons de médicaments, comme avec le système **Decagon** créé par Marinka Zitnik, Monica Agrawal et Jure Leskovec de l'Université de Stanford. La méthode utilise du deep learning de modélisation de graphes d'interactions entre molécules dans un réseau de neurones[1330]. L'IA peut aussi servir à identifier la **toxicité** de certaines molécules et de réduire la quantité de tests à réaliser sur des animaux dans les tests précliniques[1331].

Comme il faut tout de même réaliser des tests cliniques phase II et III sur des humains, on peut aussi faire appel à l'IA pour sélectionner les individus de manière à ce que leur répartition corresponde bien à ce que l'on cherche.

[1325] Le procédé utilisé par Standigm est décrit ici : http://www.standigm.com/project/.

[1326] Voir aussi Machine Learning and Synthetic Biology used to Optimize a Cellular Factory for Industrially Relevant Products par Synthiobeta, septembre 2020.

[1327] Voir How artificial intelligence is changing drug discovery de Nic Fleming dans Nature, mai 2018.

[1328] Voir Australian Researchers Have Just Released The World's First AI-Developed Vaccine par Haron Masige, juillet 2019.

[1329] Voir AI Helps Scientists Discover Powerful New Antibiotic par Mark Anderson, février 2020, en lien avec A Deep Learning Approach to Antibiotic Discovery par Jonathan M.Stokes et al, février 2020 (30 pages).

[1330] Voir Modeling polypharmacy side effects with graph convolutional networks, 2018 (9 pages et version en slides).

[1331] Voir Database analysis more reliable than animal testing for toxic chemicals du Johns Hopkins University Bloomberg School of Public Health, 2018.

C'est ce que propose **Deep 6** (2015, USA, $17M) qui utilise l'IA pour analyser des dossiers médicaux de patients candidats, notamment leurs données texte non structurées, et aider à leur sélection.

Au passage, notez que je ne cite pas de grand laboratoire pharmaceutiques. Il se trouve qu'ils communiquent peu sur leurs avancées en matière d'usages de l'IA et sous-traitent une grande partie de leur innovation amont aux startups biotechs. On n'en parle pas souvent mais ces laboratoires établis sont menacés non pas par les GAFA mais par des portefeuilles de brevets vieillissants et par une position bancale dans la chaîne de valeur. Cela se manifeste avec une rentabilité financière en baisse constante depuis une vingtaine d'années[1332].

Cancérologie

IBM a été incontestablement l'un des premiers grands acteurs à relancer la thématique de l'IA, bien avant que les GAFA s'en mêlent. La date clé à retenir est 2011 avec la victoire d'IBM Watson dans le jeu **Jeopardy** face aux meilleurs joueurs américains. Watson était au départ surtout un système de traitement du langage et de modélisation des connaissances. Aujourd'hui, c'est un ensemble de briques logicielles assemblées par les développeurs indépendants et par les équipes services d'IBM au gré des besoins des clients.

IBM Watson est décliné sur divers cas d'usages dans différents marchés verticaux. Watson n'est pas un produit plurivalent mais un ensemble de briques logicielles exploitées dans des solutions métiers développées par IBM ou ses partenaires. Il en va de même pour les briques logicielles d'IA que l'on trouve chez Google, Microsoft, Amazon et d'autres grands du logiciel et de l'Internet.

L'un des premiers marchés qu'IBM a cherché à pénétrer est celui de la santé et en particulier celui de la cancérologie. Leur ambition était d'utiliser Watson à la fois pour aider au diagnostic puis à la recommandation de traitements, le tout réalisé en partenariat avec de nombreuses cliniques américaines et dans le monde, et en exploitant les données phénotypiques (présence de la maladie, analyses biologiques et imagerie médicale) et génotypiques (variantes dans l'ADN des gènes) et les bases de connaissance du secteur composées de millions de documents (recherche, études cliniques).

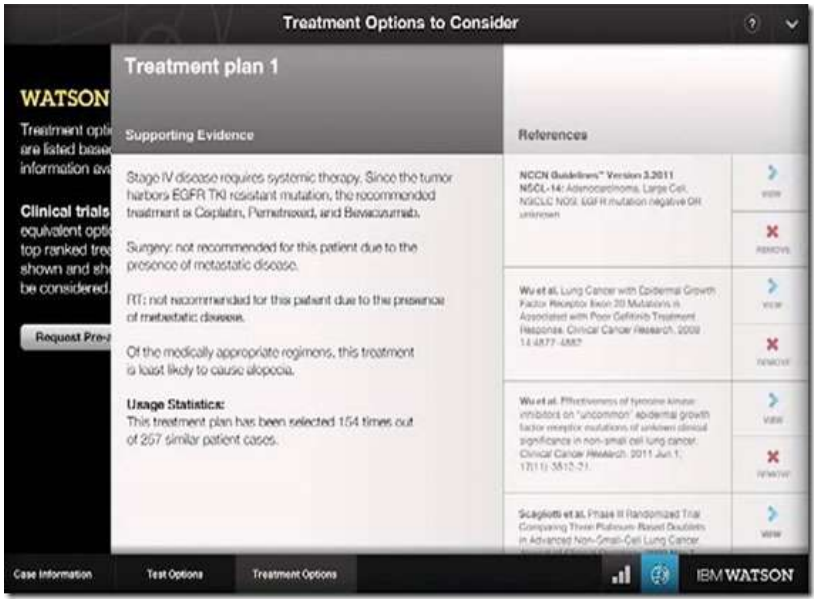
Watson for Oncology a été créé initialement en partenariat avec l'assureur santé Anthem (anciennement WellPoint) et le Memorial Sloan Kettering Cancer Center (MSK) de New York, qui associe un hôpital et un centre de recherche.

Elle a ensuite été déployée timidement dans plus d'une quinzaine d'établissements aux USA et ailleurs dans le monde comme en Inde, mais sans que l'on sache auprès de combien de praticiens et sur combien de cas de patients.

[1332] Voir Pharma's broken business model: An industry on the brink of terminal decline, de Kevin Slott, novembre 2017 qui décrit très bien les soubressauts qui affectent l'activité et le business model des industries de la pharmacie.

IBM Watson analysait les dossiers de patients atteints de tumeurs bénignes, y compris le séquençage d'ADN de cellules prélevées dans les tumeurs[1333], proposait un diagnostic, déterminait des traitements possibles et évaluait leur efficacité relative. Il devait aider à optimiser l'usage de la chirurgie, de la radiothérapie et de la chimiothérapie.

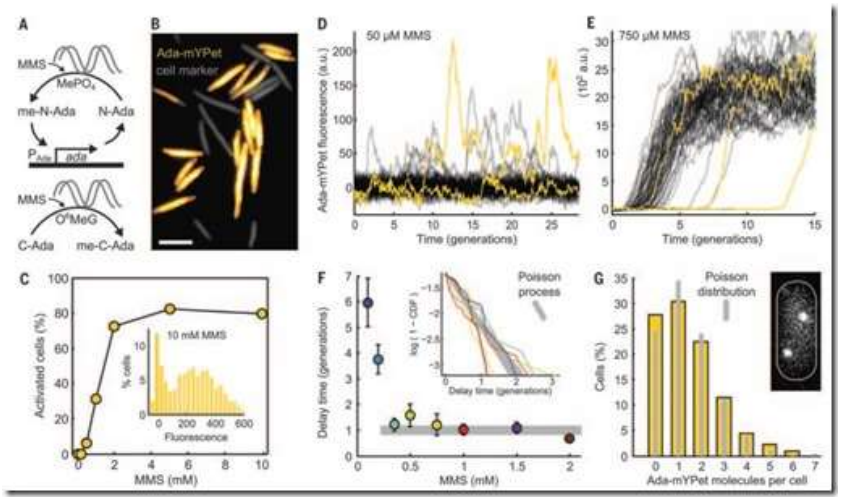
En 2014, le **Baylor College of Medicine** créait l'application KnIT (Knowledge Integration Toolkit) s'appuyant sur IBM Watson pour identifier des thérapies contre le cancer. Précisément, elle analysait la littérature scientifique pour suggérer six protéines kinases capables de contrôler le fonctionnement de la protéine p53 qui jouerait un rôle dans le développement d'environ la moitié des cancers. En 30 ans, selon IBM, moins d'une trentaine de nouvelles protéines auraient été découvertes. Ce qui mériterait d'être vérifié !



Des données statistiques peuvent exister qui font le lien entre type de thérapies et types de mutation de ces gènes. On est ici dans le domaine du big data non structuré contrairement au big data dans le marketing qui est basé sur des données bien plus structurées en général (logs Internet, données d'achats ou de consommation, bases de données relationnelles, etc). Il semble que cette partie de la solution ait été développée en partenariat avec **Cleveland Clinic**.

La solution utilise des sources d'informations variées pour faire son diagnostic et pioche notamment dans les 44 000 nouvelles publications scientifiques annuelles sur le cancer.

Les articles ne sont pas toujours faciles à exploiter : autant le texte est relativement facile à analyser, autant les illustrations qui ne sont pas toujours fournies dans un format structuré ne sont pas facilement exploitables.



Or elles fournissent des données critiques, exploitables statistiquement, à supposer que Watson puisse comprendre leur signification.

L'exploitation de la littérature scientifique ne doit donc pas être bien évidente à ce niveau. Par contre, elle est peut-être plus aisée pour les études liées aux AMM (autorisations de mise sur le marché) et autres études épidémiologiques.

[1333] Semble-t-il, et non pas un simple génotypage, mais on peut aussi séquencer l'ARN qui évalue l'expression des gènes dans les tumeurs.

On se demande par contre si ce genre de solution sait tenir compte de la forte proportion de publications scientifiques qui est entachée de fraudes ou exagérations[1334].

Dans les démonstrations d'IBM, la solution de cancérologie à base de Watson fournissait au praticien un choix de traitements qui sont fournis avec un indice de confiance, comme la probabilité de survie. Après avoir démarré avec les cancers du poumon, étaient ajoutés les leucémies, les mélanomes, les cancers du pancréas, des ovaires, du cerveau, du sein et du colon.

Dans cette application, Watson était censé battre l'homme dans la force brute : il compulsait notamment des bases de données de recherche en oncologie pour aider les cancérologues. Mais il ne faisait pas directement progresser la recherche sur les cancers. Les articles scientifiques exploités ont chacun nécessité de 3 à 7 années de recherche réalisées par plusieurs chercheurs ! C'est un travail considérable. Watson utilisait donc les résultats de la recherche existante qui s'appuie sur des expériences in-vitro et in-vivo, que l'on ne sait pas encore simuler numériquement, et les résultats statistiques associés. Bref, on a encore besoin de chercheurs ! Pour automatiser ce processus, il faudrait passer par plusieurs stades d'évolution de l'IA : ajouter la dimension créative et conceptuelle, automatiser des tests in-vitro et in-vivo avec des robots et en dernier lieu, bien plus tard, réaliser ces tests in-silico quand les algorithmes et la puissance de calcul le permettront.

Tout cela est bien merveilleux mais le marketing d'IBM autour du cancer était trompeur et ses prouesses largement survendues. Les annonces évoquées ci-dessus n'ont jamais été éprouvées ni déployées à grande échelle[1335]. En 2018, on pouvait même constater une grosse déconvenue sur le sujet même si IBM avait du mal à l'admettre[1336]. En août 2018, IBM contre-attaquait en indiquant que Watson fournissait une aide appréciable aux cancérologues[1337]. Watson for Oncology serait utilisé dans 230 hôpitaux avec 84 000 patients traités sur H1 2018.

Il fallut attendre début 2019 pour qu'IBM jette finalement l'éponge en cancérologie, citant de « mauvaises ventes »[1338]. Les mauvaises nouvelles s'amoncelaient avec des rapports selon lesquels IBM recommandait les mauvais traitements[1339]. Ils stoppaient également l'activité IBM Watson en drug discovery avec les entreprises de pharmacie.

[1334] Dans How to Make More Published Research True, John Ioannidis indiquait en 2014 que 85% des ressources des chercheurs sont gaspillées et leurs résultats publiés sont faux ou exagérés. Dans Raise standards for preclinical cancer research, Glenn Begley et Lee Ellis indiquaient en 2012 que 90% des résultats de 53 études majeures dans le domaine du cancer n'étaient pas reproductibles. Donc, si elles sont utilisées par Watson, il ne peut pas en sortir grand-chose d'utile ! Voire, cela peut même être dangereux. Enfin, dans Believe it or not: how much can we rely on published data on potential drug targets?, Florian Prinz, Thomas Schlange et Khusru Asadullah indiquaient en 2011 que 79% des résultats de 67 travaux de recherche en cancérologie et cardiologie n'étaient pas reproductibles chez Bayer. Qui plus est, les recherches qui donnent lieu à des résultats négatifs sont bien moins publiées que celles qui sont concluantes. Ce sont toutes ces études qui alimentent Watson for Oncology ! Le biais statistique qu'elles induisent est énorme ! Source de cette liste : Bio-Modeling Systems - The Mechanisms-Based Medicine Company de Manuel Gea, juillet 2017. D'où l'intérêt d'initiatives comme le centre METRICS de Stanford, qui vise à faire de la méta-recherche, donc d'auditer les pratiques des chercheurs pour les améliorer.

[1335] Voir IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close de Casey Ross., septembre 2017 et Why Everyone Is Hating on IBM Watson—Including the People Who Helped Make It de Jennings Brown en octobre 2017.

[1336] IBM's Watson reportedly created unsafe cancer treatment plans, juillet 2018, et Report: IBM Watson delivered 'unsafe and inaccurate' cancer recommendations, juin 2018 fait état de documents internes de l'équipe IBM Watson Health qui indiqueraient que les résultats de Watson for Oncology comprendraient des résultats incorrects de propositions de thérapies dans sa mise en œuvre sur des cas fictifs portant sur 13 typs de cancers au Memorial Sloan Kettering Cancer Center. Ce qui n'est pas étonnant quand on sait qu'il exploite des méthodes probabilistes qui ne peuvent garantir 100% de bons résultats. IBM licenciait en mai 2018 des équipes issues de trois acquisitions dans la santé (**Explorys**, **Truven** et jusqu'à 80% des salariés de **Phytel**) acquises en 2015 et 2016. Ils ont surtout des difficultés économiques à faire sortir l'IA des PoC (proofs of concepts). Certaines sont liées aux limitations des solutions, d'autres aux difficultés classiques de déploiements d'innovations. Et, avec ou sans IA, il est fréquent que des acquisitions de startups par de grands acteurs finissent mal. Voir Layoffs at Watson Health Reveal IBM's Problem With AI, juin 2018.

[1337] Voir IBM pushes back on negative Watson Health stories, août 2018.

[1338] Voir IBM Halts Sales of Watson AI For Drug Discovery and Research par Joel Hruska, avril 2019, qui fait reference à How IBM Watson Overpromised and Underdelivered on AI Health Care par Eliza Strickland, avril 2019.

[1339] Voir IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show par Casey Works, juillet 2018.

Il leur restait notamment une activité en génomique avec **IBM Watson for Genomics** qui est utilisable… en cancérologie[1340] (vidéo) ainsi que le traitement de l'image exploité en imagerie médicale.

Faute de pouvoir en faire un véritable business, des chercheurs d'**IBM** Zurich ont toutefois diffusé trois solutions de cancérologie en open source et dans le domaine public[1341] qui servent à comprendre les mécanismes chimiques déclencheurs de maladies plurifactorielles comme les cancers. PaccMann (Prediction of anticancer compound sensitivity with Multi-modal attention-based neural networks) utilise un réseau de neurones multimodal pour prévoir l'efficacité de traitements à partir de données sur l'expression des gênes et sur la structure moléculaire des traitements, INtERAcT (Interaction Network infErence from vectoR representATions of words) extrait les données de publications scientifiques dans un domaine précis, les interactions entre protéines et PIMKL (pathway-induced multiple kernel learning) prédit la progression de cancers.

Enfin, dans le cadre de sa restructuration en deux sociétés annoncée fin 2020, IBM semblait vouloir céder son activité IBM Watson Health selon des informations diffusées début 2021.

Mais une startup américaine semble s'être lancée sur un positionnement voisin de celui qu'IBM a abandonné. **Tempus** (2015, USA, $520M) combine ainsi les séquençages génétiques de tumeurs cancéreuses, des données diverses (rapports, recherche, …), de l'imagerie médicale et du machine learning pour proposer des traitements personnalisés aux patients. J'espère pour eux et pour les patients cobayes qu'il ne s'agit pas d'un nouveau Theranos !

Heureusement, il n'y a pas qu'IBM Watson dans les usages de l'IA en cancérologie. De nombreux chercheurs investiguent des méthodes innovantes pour améliorer les traitements. Mais leurs résultats sont moins époustouflants car ils concernent des cas très précis et n'ont pas des prétentions généralistes comme IBM avec Watson.

C'est le cas de cette nouvelle méthode de traitement des glioblastomes, ces tumeurs cérébrales très difficiles à traiter, avec un ciblage précis des chimiothérapies via un système à base d'apprentissage par renforcement qui optimise à la fois l'efficacité du traitement et minimise le nombre de doses. Le modèle créé par des chercheurs du **MIT Media Lab** a été entraîné avec des tests cliniques sur 50 patients et 20 000 tests pour alimenter l'apprentissage par renforcement. Le modèle entraîné a été ensuite testé sur 50 autres patients "simulés". Reste à le faire sur de véritables patients[1342] !

Des méthodes équivalentes sont développées pour optimiser les radiothérapies. C'est le cas de travaux de l'**Université de Toronto**[1343] qui raccourcissent le temps de préparation des séances. Ils utilisent du machine learning à base de PCA (principal component analysis, pour déterminer des variables clés) et de régressions pour analyser un historique de données de radiothérapies et créer des plans de séances optimisés. Le système a été évalué sur 217 patients atteints de cancers de la tête et du cou. La startup **TheraPanacea** (2017, France) est positionnée sur un créneau voisin.

Turbine (2015, Hongrie) développe des outils de simulation logiciels permettant de créer des thérapies anti-cancéreuses. Ils créent des doubles numériques de cellules cancéreuses pour tester ces thérapies. Ils simulent les mécanismes de l'apoptose (mort programmée de la cellule), leur prolifération et les activités métastatiques. Les simulations sont conduites avec des millions de combinaisons de traitements[1344].

[1340] Voir IBM's Watson for Genomics Launches in First European Hospital par Clara Rodriguez Fernandez, septembre 2019.

[1341] Voir IBM gives cancer-killing drug AI project to the open source community par Charlie Osborne, juillet 2019.

[1342] Voir Reinforcement learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regiment Selection de Pratik Shah et Gregory Yauney du MIT Media Lab, 2018 (65 pages).

[1343] Voir Knowledge - based automated planning for oropharyngeal cancer, 2017 (67 pages).

[1344] Voir Simulating Cancer Cells Using Artificial Intelligence, Nanalyze, août 2019.

Des chercheurs de **Cincinnati** ont créé une IA qui aide à trouver les bons volontaires pour des traitements en essais clinique à partir de leurs dossiers cliniques de services d'urgences. L'ACTES (Automated Clinical Trial Eligibility Screener) permet de réduire le temps de sélection des patients de 34% et d'améliorer ladite sélection.

Le système utilise du machine learning de données structurées et du traitement du langage[1345]. Dans la même veine, une IA servait à sélectionner des femmes atteintes de **cancers des ovaires** pouvant tester des immunothérapies ciblées[1346]. Autre approche, évaluée par la Case Western Reserve University de Cleveland dans l'Ohio, l'analyse de corrélation entre scanner du poumon et l'efficacité d'immunothérapies de traitement du cancer du poumon[1347].

DeepLife.co (2019, France) exploite le deep learning pour faire du reciblage thérapeutique ou créer de nouvelles molécules thérapeutiques.

Enfin, **Eurekam** (2012, France, 1,5M€) propose Drugcam, un système de contrôle du dosage des préparations de chimiothérapie par analyse de vidéos qui limite les erreurs médicamenteuses.

Chirurgie

La chirurgie fait évidemment de plus en plus appel à des robots. Nombre d'entre eux répètent des gestes de chirurgiens à distance comme les Da Vinci d'**Intuitive surgical** (1999, USA) qui sont spécialisés dans les opérations de l'abdomen et sont déployés depuis plusieurs années. Seuls certains ont une véritable IA lorsqu'ils sont autonomes et exploitent par exemple le résultat d'imagerie médicale réalisée en temps réel sur les patients.

L'IA peut aussi servir à former les chirurgiens. C'est ce que propose **Insimo** (2013, France) qui développe un simulateur réaliste d'organes à base d'IA et d'outils de simulation.

De manière encore plus avancée, **Cambridge Bio-Augmentation Systems** (2015, UK) utilise du machine learning pour comprendre le fonctionnement des nerfs de patients pour leur greffer des membres bioniques. Leur Prosthetic Interface Device (PID) est adapté à l'ajout d'un bras artificiel.

Enfin, les anesthésistes pourront aussi bientôt exploiter l'IA pour mieux anticiper les complications des patients, notamment les risques d'hypoxémie, en fonction des nombreux paramètres qui les caractérisent. C'est l'objet du projet de recherche Prescience de l'**Université de Washington** (dans l'État du même nom, à Seattle)[1348].

Diabète

Les solutions aidant les diabétiques à suivre leur traitement et à équilibrer leur dose d'insuline (pour le type 1) et leur alimentation sont très nombreuses et font appel au machine learning.

Diabeloop (2015, France, 47,3M€) finalise la mise au point de DBLG1, une solution complète de bout en bout, intégrant un capteur de glycémie G6 en continu d'origine Dexcom, une pompe à insuline et un boîtier dédié faisant tourner un logiciel de suivi pour ajuster automatiquement le dosage d'insuline, exploitant des briques d'IA. Le DBLG1 a obtenu l'agrément CE en novembre 2018, ouvrant la voie à une commercialisation prochaine. Ce n'est pas la seule solution de gestion du cycle glycémique en boucle mais semble être la seule qui exploite l'IA.

[1345] Voir Artificial intelligence solution improves clinical trial recruitment, 2019.

[1346] Voir AI Picks Out 'Shapeshifting' Cancer Cells, Revealing Potential New Drug Targets par Victoria Forster, 2019.

[1347] Voir AI Tool to Predict Checkpoint Therapy's Effectiveness par Siavash Parkhideh, décembre 2019.

[1348] Voir Machine-Learning Can Help Anesthesiologists Foresee Complications de Marc O'Reilly, octobre 2018 ainsi qu'une version préliminaire de leur publication dans Anesthesiologist-level forecasting of hypoxemia with only SpO2 data using deep learning, décembre 2017 (6 pages). La méthode associe des Convnets et des réseaux à mémoire de type LSTM.

Beta Bionics (2015, USA, $132M) a l'air d'être positionné sur le même créneau pour créer un apareil de régulation automatique de la glycémie asservi par IA. La pompe à insuline qui comprend visiblement aussi un système de suivi en continu de la glycémie dans le sang (à vérifier) délivre une dose régulière calculée avec du machine learning. La complexité vient du fait qu'il faut tenir compte de l'effet retard de l'effet de l'insuline comme de celui de l'alimentation ou de l'activité physique.

Hillo (2016, France, 2,1M€), anciennement Healsy se focalise pour sa part sur l'algorithme de régulation de la prise d'insuline construit aussi à partir de modèles de machine learning. **Medtronic** (USA) a développé SUGAR.IQ, une solution du même genre qui permet de prédire sa glycémie en fonction de divers paramètres dont l'activité physique et les prises de repas.

Cognitive Scale (2013, USA, $40M) a créé la solution Cognitive Clouds qui est proposée aux adolescents atteints de diabète type 1 pour les aider à se réguler, en intégrant les aspects médicaux (prise d'insuline, suivi de glycémie), d'activité physique et d'alimentation. Il y a des dizaines de startups qui visent le même marché et avec plus ou moins de bonheur. Très souvent, elles méconnaissent le fonctionnement des diabétiques dans la régulation de leur vie et leur segmentation.

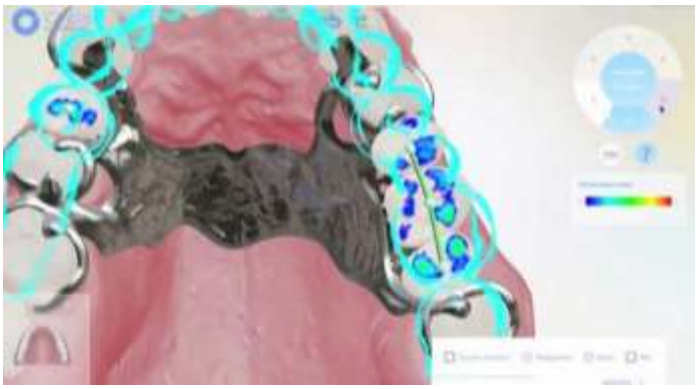
Diabnext (2018, France) est une startup à cheval sur la France, les USA à Boston et Taïwan qui est positionnée sur le même créneau que Cognitive Scale.

Orthodontie

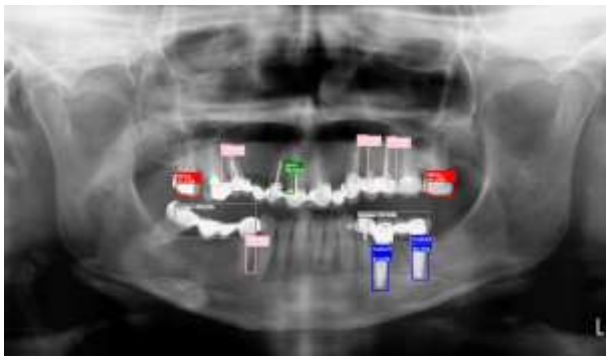
Les traitements dentaires peuvent faire appel à de l'IA pour l'analyse de radios, la simulation d'implants et en robotique chirurgicale.

Biotech Dental (1987, France) a lancé en mars 2018 Lucy, un logiciel de simulation biomécanique pour la création de prothèses dentaires. C'est plutôt un logiciel de modélisation et de CAO que d'IA. Mais il propose des traitements qui sont issus d'un entraînement exploitant des études de cas précises. Il s'appuie visiblement sur les travaux d'**Anatoscope** (2015, France) qui fait de la simulation 3D (vidéo). Les prothèses sont ensuite imprimées en 3D (vidéo).

Complémentaire à la solution de Biotech Dental, **Neocis** (2012, USA, $20,9M) a développé le robot de pose d'implants dentaires Yomi. Vous êtes partants pour le tester ? Pour vous rassurer, il est piloté par le dentiste et ne fonctionne pas en roue libre. Reste à savoir si ce robot contient véritablement des briques d'IA !



Dentem (2016, Canada) utilise la vision artificielle pour analyser les radios dentaires panoramiques et identifier les problèmes avec la solution Dx Vision. Il cartographie automatiquement les implants, amalgames, couronnes et caries.

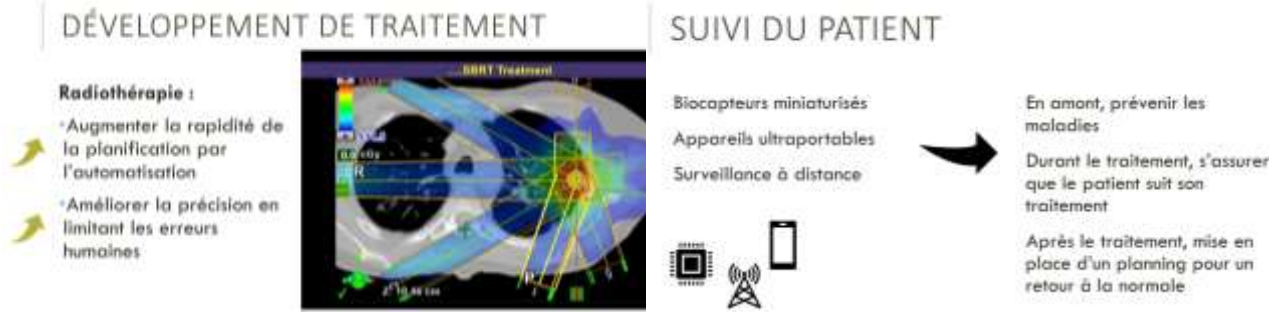


La startup propose aussi un logiciel de gestion de rendez-vous et de gestion de dossier patient, deux fonctions assez classiques dans ce marché.

Neuro et psy

L'IA a aussi quelques domaines d'applications dans le traitement de pathologies neurologiques et psychiatriques.

Intradys (France) utilise le machine learning radiologie préopératoire pour préparer des interventions sur le cerveau consécutives à des AVC. Leur solution permet de traiter avec plus de précision des anévrismes aussi bien en chirurgie qu'en radiologie ciblée.



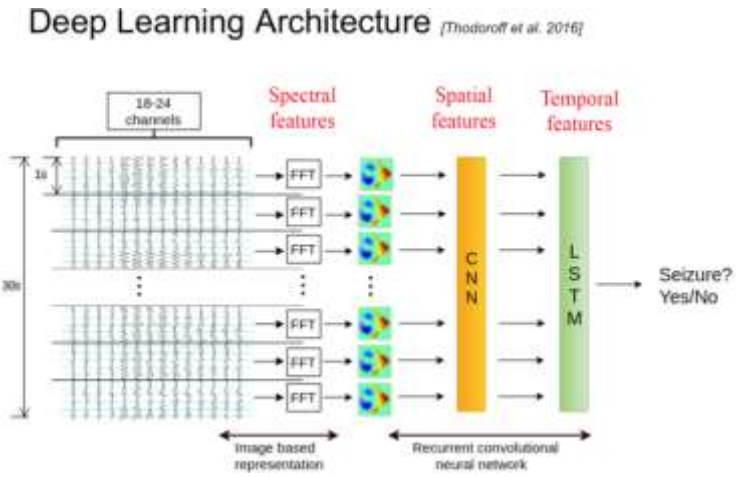
Textpert (2015, USA) a développé son « Empathetic Virtual Interface » avec des chercheurs d'UCLA et de l'USC. Il s'agit d'un chatbot de diagnostic psychothérapeute. Il devrait être au point d'ici fin 2019. Mais l'histoire ne dit pas s'il passera le test de Turing à cette échéance. Il ne coûtera que $10 par mois, soit beaucoup moins qu'un psychothérapeute, en tout cas aux USA.

Dans un projet de recherche européen Horizon 2020 anglo-japonais, le robot **Nao** peut former les enfants autistes à bien reconnaître les émotions (vidéo)[1349].

En effet miroir, **Emoface** (France) permet à un utilisateur d'animer un avatar 3D expressif. La solution est actuellement testée par des enfants autistes pour les aider à apprendre à émettre des émotions.

[1349] Vu dans Humanoid Robot Teaches Autistic Kids to Recognize Emotions, juillet 2018.

La détection d'épilepsie fait partie des travaux du FAIR, le laboratoire de recherche en IA de **Facebook**, avec de l'apprentissage par renforcement qui associe une analyse spectrale (CNN) et temporelle (LSTM) d'électro-encéphalogramme[1350]. Le projet est piloté par Joëlle Pineau, qui dirige la branche de Montréal de ce laboratoire. Le principe consiste à générer une neurostimulation artificielle qui évite le déclenchement inopiné d'épilepsie chez les patients.



Un projet équivalent de diagnostic de l'épilepsie était publié en 2020 et réalisé par une équipe australienne. Il s'appuyait sur l'analyse de (seulement) 400 enregistrements d'EEG de patients sains ou atteints d'épilepsie. L'outil développé pourrait notamment servir à la formation des neurologues[1351]. En 2020, l'université de Louisiane annonçait une solution permettant de prédire une heure à l'avance les crises d'épilepsie avec une précision de 99,6%[1352]. La solution combine quatre réseaux de neurones dont un CONVNET et un RNN.

Il existe enfin plusieurs initiatives destinées à gérer la reconnaissance automatique du langage des gestes des sourds et malentendants. Les techniques utilisées combinent souvent plusieurs modèles : un réseau convolutif pour reconnaître la position des mains et du corps, puis des modèles pour détecter l'évolution de la position des mains dans le temps (réseau à mémoire ou modèles de Markov). Ce sont pour l'instant surtout des projets universitaires ou de recherche[1353].

Exosquelettes

De leur côté, les exosquelettes ont moins de capteurs sensoriels. Ils sont surtout pilotés par l'utilisateur, notamment via la partie de leur corps qui fonctionne encore comme pour **Wandercraft** (2013, France, $30,5M), qui a démarré la commercialisation de sa solution en 2019 après de nombreux tests probants, ou **Cyberdyne** (2004, Japon).

Ce genre de solution peut exploiter des électrodes placées sur le cortex moteur dans les lobes biparietaux comme pour le projet d'exosquelette à quatre membres du laboratoire grenoblois **Clinatec** qui est destiné aux tétraplégiques. Après des années de mise au point (je les avais découverts en 2015 et le laboratoire avait été créé en 2011), il était testé pour la première fois avec un patient en octobre 2019 (vidéo).

[1350] Vu dans Adaptive treatment of epilepsy via reinforcement learning de Joëlle Pineau, 2017 (35 slides).

[1351] Voir Monash University researchers speed up epilepsy diagnosis with machine learning par Aimee Chanthadavong, septembre 2020.

[1352] Voir New AI System Predicts Seizures With Near-Perfect Accuracy par Michelle Hampson, novembre 2019.

[1353] Voir par exemple How we used AI to translate sign language in real time, septembre 2018, un projet australien.

Ekso Bionics (2005, USA, $70,8M) vise de son côté aussi bien les compensations de handicaps que les applications industrielles. Ses exosquelettes sont en tests dans des usines de Ford depuis fin 2017[1354]. **RB3D** (2001, France) met au point pour sa part des exosquelettes ciblant notamment le marché de la construction.

Il y a beaucoup de mécanique dans ces produits avec des enjeux de miniaturisation des moteurs et des batteries pour les rendre aussi légers et pratiques que possible. Les exosquelettes peuvent être partiels comme avec la main robotisée de **BrainRobotics** (2015, USA) et l'exosquelette lombaire Atlas de **Japet Medical Devices** (2016, France, $1,7M).

De plus en plus, on fait appel à de l'IA pour ajuster leur fonctionnement et le rendre aussi naturel que possible pour leurs utilisateurs. Les Universités de Caroline du Nord et d'Arizona ont mis au point une prothèse de genoux qui adapte son fonctionnement rapidement – en une dizaine de minutes - grâce à une IA embarquée utilisant de l'apprentissage par renforcement. Elle évite un long processus de rééducation[1355]. Des chercheurs de l'EPFL en Suisse ont fait de même pour accélérer la prise en main de prothèses de membres supérieurs, dans un premier temps avec un exorobot[1356].

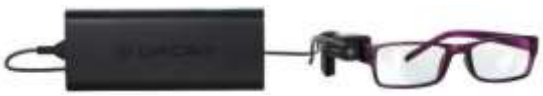
Hoobox Robotics (Brésil) a créé la chaise roulante Wheelie 7 qui est contrôlée par la reconnaissance des expressions sur le visage de leurs utilisateurs qui peuvent être atteints de divers troubles neurobiologiques. Ils s'appuient sur la technologie et le capteur 3D RealSense d'Intel (vidéo).

Ophtalmologie

Panda Guide (2015, France) propose un système dédié aux aveugles qui se positionne autour du cou et est complété d'oreillettes audio. La partie IA tourne dans le smartphone en mode offline avec un modèle de vision entraîné sur serveur, capable de reconnaître un millier d'objets de la vie courante. **Microsoft** propose une solution qui a l'air d'être voisine, SeeingAI, et qui fonctionne dans 70 langues.

OrCam Technologies (2010, Israël, $86,4M) apporte la vision aux mal voyant via une caméra reliée à un système de reconnaissance d'objets qui décrit les scènes de manière vocale. C'est une startup israélienne. On a ici un mélange de computer vision et de text to speech. Ils en proposent une déclinaison pour dyslexiques sous la forme d'un stylo réalisant de la reconnaissance d'écriture sur du texte imprimé, l'Orcam Read

La caméra miniature se positionne sur des lunettes traditionnelles et comprend un écouteur.



[1354] Voir Ford Testing Exoskeleton Suit For Its Factory Workers, novembre 2017.

[1355] Voir Online Reinforcement Learning Control for the Personalization of a Robotic Knee Prosthesis, IEEE, janvier 2019.

[1356] Voir This prosthetic arm combines manual control with machine learning par Devin Coldewey, septembre 2019.

Systèmes de santé

L'irruption de l'IA dans les systèmes de santé n'est pas anodine. Elle peut les transformer radicalement, faire évoluer de nombreux métiers, amener à revoir certaines organisations, notamment dans les hôpitaux et modifier la relation avec les patients[1357].

Les solutions de systèmes de santé couvrent des besoins divers : le suivi de l'observance des traitements, l'évitement d'erreurs de prises de médicaments, les robots pour s'occuper des personnes âgées, le suivi des dépenses de santé ou l'optimisation des ressources des hôpitaux et praticiens. Ils génèrent de gros volume de données, d'où les nombreux cas d'usage potentiels de l'IA.

La mise en place de toutes ces solutions va entraîner son lot de questionnements et de réglementation. Il sera dans certains cas associés à la création d'IA explicables permettant d'évaluer leurs diagnostics ou préconisations[1358].

En voici quelques exemples, toujours pris dans l'univers florissant des startups.

MedAware (2012, Israël, $10,4M) fournit une solution qui permet d'éviter les erreurs de prescription médicamenteuse en temps réel pour les médecins. Avec des morceaux de big data et de machine learning dedans qui exploitent notamment des bases de données médicales d'historiques de patients.

Aux USA, le projet **Deep Patient** peut prédire l'apparition de 78 pathologies via un entraînement exploitant les dossiers médicaux de 1,6 millions de patients de l'hôpital MtSinai de New York étalés sur plusieurs décennies[1359]. Il a entre autres bénéficié de la loi *Health Information Technology for Economic and Clinical Health Act* votée en 2009, juste avant l'Affordable Care Act (Obamacare). Elle encourageait les hôpitaux et médecins à adopter des systèmes de dossiers médicaux interopérables. 84% des hôpitaux avaient emboîté le pas en 2017.

ExactCure (2018, France, 1M€) propose un jumeau numérique du patient qui permet de simuler les effets des médicaments dans leur corps en fonction de leurs caractéristiques (âge, poids, statut rénal, génotype, tabagisme, diabète, etc). Cela sert notamment à évider les surdosages ou sous-dosages de médicaments ou des contre-indications liés à des interactions entre traitements. Le machine learning est exploité pour identifier des corrélations entre situations et dosages ou contre-indications.

Predical (2015, France) fait partie de ces startups de la silver économie et des objets connectés qui veut faire du prédictif avec de l'apprentissage d'IA par la détection des routines de la vie et les anomalies associées.

Somatix (Israël) a développé un système de détection des gestes via des wearables. Leurs clients basés aux USA sont des hôpitaux. La solution disponible sur le marché est SmokeBeat. Elle permet une télésurveillance des patients qui suivent un traitement anti-tabac. Ils développent aussi une solution de suivi des soins des seniors.

macro-eyes (2014, USA) propose un outil original qui permet de mieux gérer la prise de rendez-vous avec les médecins tout en réduisant les risques d'absence des patients.

CLinigrid (2005, France) est un créateur de solutions d'exploitation des données de santé qui s'appuie notamment sur du machine learning pour identifier des corrélations diverses.

[1357] Voir Transforming healthcare with AI - The impact on the workforce and organisations, McKinsey pour EIT Health, mars 2020 (134 pages).

[1358] Voir à ce sujet l'initiative **Ethik-IA** qui est dédiée au monde de la santé, lancée en 2017 par David Grison, l'ex-délégué général de la Fédération hospitalière de France (FHF). Elle se présente comme une initiative académique et citoyenne qui veut défendre une régulation positive de l'IA et de la robotisation dans la santé.

[1359] Voir Deep Patient: Predict the Medical Future of Patients with Artificial Intelligence and EHRs de Riccardo Miotto, 2017 (51 slides) ainsi que Dans Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis, février 2018 (16 pages) qui fait le point des méthodes à base d'IA d'exploitation des dossiers médicaux aux USA.

En France, Emmanuel Bacry (CNRS) exploite les données des parcours de santé de la CNAM pour d'identifier des effets secondaires indésirables de médicaments comme pour l'antidiabétique oral Pioglitazone qui a été retiré du marché en 2011 à cause de son incidence sur le cancer de la vessie[1360]. Les données utilisées proviennent du SNIIRAM (Système National d'Information Inter-régimes de l'Assurance-maladie) qui comprend un milliard de feuilles de soins de 65 millions d'assurés.

NeedaBot (2016, France), anciennement HoCaRo, propose un robot de soutien aux aides-soignants dans les EHPAD. Avec bras, tête et roulettes. Il permet d'éviter et de détecter les chutes.

Diligent Robotics (2017, USA, $5,8M) propose son robot Moxi qui assiste les personnels soignants dans les hôpitaux pour le transport d'objets divers et de médicaments. Il est en tests depuis 2019 dans quatre hôpitaux américains, au Texas. Il dispose d'un bras articulé, d'un buste posé sur un système à roulettes. La mécanique n'a donc rien de particulièrement original (vidéo).

Events Bots (2016, France) a développé Hope, un robot d'accompagnement d'enfants malades en milieu hospitalier, qu'il occupe pendant les soins.

Hindsait (2013, USA) a une solution en cloud servant à identifier les déviations dans les dépenses de santé. Cela sert donc surtout aux financeurs des systèmes de santé que sont les assurances publiques, privées et les mutuelles. Ca fait moins rêver le patient ! Ce genre de solution fait du prédictif multicritère à base de machine learning.

Senscio Systems (2009, USA, $7,9M) commercialise Ibis Health Management Solution qui est censé réduire les coûts de santé pour des organismes payeurs en améliorant le suivi de maladies chroniques.

Le principe ? Détecter les déviations des patients par rapport aux prescriptions des médecins. Cela a l'air de supposer que le patient utilise un grand nombre d'objets connectés pour suivre sa tension, sa glycémie, son activité physique et son poids, entre autres paramètres vitaux. **Innovaccer** (2014, USA, $43M) semble être sur le même marché.

L'IA peut aussi servir de manière plus générique à faciliter le travail des médecins. C'est ce que propose **Suki.ai** (2016, USA, $40M) avec son système de dictée de notes vocales permettant de gagner du temps dans la rédaction du dossier de santé des patients. Cela permet de diviser par dix le temps alloué à cette tâche[1361]. Mais en anglais seulement j'ai l'impression.

Finance

Friand de données, le secteur de la finance est un terrain très favorable à l'usage de l'IA et dans tous ses métiers et recoins[1362], du front office au back office en passant par la relation clients via des chatbots, l'analyse de risques et l'optimisation de portefeuille, sans compter les techniques de base utilisées depuis longtemps comme la reconnaissance automatique de l'écriture manuscrite dans les chèques.

[1360] Voir Big data : premier succès dans l'alerte sanitaire, janvier 2018.

[1361] Voir This AI Company Is Trying To Fix Physician Burnout - An Interview With Dr. Nathan Gunn, COO Of Suki.ai par Jun Wu, 2020.

[1362] Le schéma vient de Tomorrow's AI-Enabled Banking de IPSoft.

L'objectif est toujours d'optimiser les opérations[1363], de maximiser les rendements financiers, d'en réduire les coûts, de personnaliser les offres et d'améliorer la relation client. Les services financiers donnent lieu à la création d'un grand nombre de startups qui fournissent quelques indications des usages prometteurs de l'IA dans les métiers de la finance.

Le marché bancaire est très verticalisé et a plutôt bien résisté aux coups de boutoir des startups depuis plus de 20 ans, malgré une insatisfaction chronique de certains segments de clients. La banque directe se développe lentement, surtout en France.

Avec ou sans IA, les Fintechs ambitionnent de disrupter le marché à tous les étages (mobilité, Bitcoins, crowdfunding, etc) mais sont encore très focalisées sur les moyens de paiement et moins sur la gestion de comptes et la commercialisation de produits financiers.

La banque de détail a déjà traversé une partie de sa révolution numérique en transférant une part du travail salarié des agences et du back-office vers les clients se débrouillant par eux-mêmes sur les guichets automatiques et autres applications web et mobiles. Cela les a amenés à revoir le métier des agences, plus positionnées sur la commercialisation de produits financiers. Cette transformation correspond assez bien à ce qui se produira dans les années qui viennent dans divers métiers de services comme dans l'expertise comptable.

L'IA pourrait-elle accélérer la mutation du secteur bancaire ? Est-elle un facteur qui peut faire évoluer la notion de confiance, critique dans le secteur, d'un type d'acteur à un autre [1364]?

La question reste ouverte. Si les banques commencent à se restructurer, ce n'est d'ailleurs pas à cause de l'IA mais plutôt de l'émergence de nouveaux acteurs et des nouvelles habitudes numériques des clients.

L'inquiétude est cependant grandissante de voir les applications de l'IA accélérer la décrue côté emplois dans le secteur, aussi bien dans la banque de détail que pour les métiers les mieux payés dans les banques d'investissements[1365].

Optimisation d'investissements

L'exploitation de l'IA couvre toutes les solutions d'optimisation de gestion des investissements, surtout boursiers. La majorité des solutions du marché côté startups sont b2c et quelques-unes sont b2b. Les solutions b2c ciblent en premier lieu les audiences adultes les plus jeunes, ces fameux millenials.

[1363] Cela peut aller jusqu'à prévoir le besoin en cash pour les distributeurs automatiques de billet.

[1364] Le problème étant que le grand public n'a plus confiance en grand-chose. Tout le monde en prend pour son grade : les politiques, les médias et les entreprises. Voir le baromètre mondial de la confiance 2020 d'Edelman.

[1365] Voir L'IA pourrait remplacer près de 6 millions d'emplois dans la finance par Robin Tutenges, décembre 2019 et Robots in Finance Could Wipe Out Some of Its Highest-Paying Jobs par Lananh Nguyen, décembre 2019.

L'optimisation d'investissements boursiers est proposée par **AdvisorEngine** (2014, USA, $54,3M) avec son logiciel en cloud de gestion d'investissements personnels pour les conseillers en gestion de patrimoine et leurs clients (*ci-contre*). Ils utilisent du machine learning mais ne précisent pas comment dans leur communication comme nombre de startups de ce secteur, ni quelles données servent à l'entraînement des IA.

WealthArc (2015, USA, $1,5M) propose une solution voisine également destinée aux conseils en gestion de patrimoine (vidéos de démonstration).

Acorns (2012, USA, $152M) automatise la création de portefeuilles financiers selon le niveau de risque souhaité. Ils ont associé leur offre à un système de cashback avec quelques retailers et marques comme Apple, Airbnb, DirecTV. Cela permet indirectement de faire des économies à placer à l'occasion des dépenses associées.

Stash (2015, USA, $189M) permet de gérer ses investissements personnels de manière sélective en privilégiant par exemple les sociétés qui protègent l'environnement. La startup gère aussi vos comptes en banque et votre carte de crédit.

La gestion financière en ligne est proposée par **Wealthfront** (2008, USA, $204M) et **Betterment** (2008, USA, $275M), deux startups extrêmement bien financées et qui ne ciblent visiblement que le marché US. La première a une offre orientée « scénarios » qui permet de planifier le financement des événements de la vie (achat logement, voyages, études des enfants, retraite…).

Pefin (2015, USA) gère aussi la finance personnelle des foyers de tous niveaux de patrimoines, aussi bien pour planifier les investissements que pour gérer ses comptes au plus juste, selon les situations. La startup indique surveiller 2 millions de données pour optimiser ses recommandations : dépenses personnelles, indicateurs économiques divers, fiscalité, couverture santé, etc (vidéo).

WealthInitiative (2015, Suisse) aide les gestionnaires de fortune à conseiller leurs clients, notamment pour les placements dans l'immobilier, l'art et autres objets de luxe (avions, voitures de collection, bateaux, pièces de monnaie, bijoux, timbres de collection…).

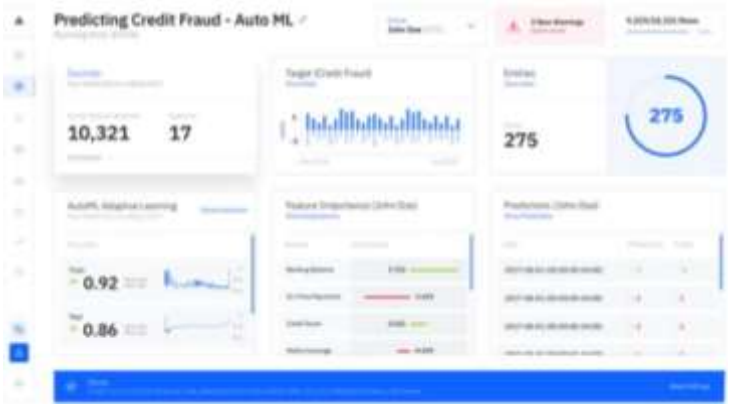
Des algorithmes à base de machine learning servent à évaluer ces différents biens. La startup souhaite surtout devenir un nouvel intermédiaire pour la vente de ces biens entre grandes fortunes. Cette activité permet de collecter des valeurs de transactions et d'alimenter le modèle de machine learning de la solution, en plus de générer des commissions de ventes.

De nombreux fonds d'investissement et fournisseurs de logiciels dans ce marché essayent de tirer parti de l'IA pour prévoir l'évolution des cours et optimiser leurs stratégies d'investissement. Très souvent, ce sont les fonds eux-mêmes qui créent leurs propres solutions logicielles. C'est une question de logique : il n'y aurait pas d'asymétrie de marché entre les gagnants et les perdants si tous les investisseurs utilisaient une même IA « nivelante » pour gérer leurs stratégies d'investissement. Un cas typique est incarné par le fonds d'investissement spéculatif **J4Capital** (2017, USA) lancé par un certain Jeff Glickman. Il utilise une "superintelligence" qui a l'air un peu magique permettant de prédire l'évolution de la bourse « *avec 60% de réussite* ». Cela semble un peu flou et bien survendu, surtout dans la mesure où cette performance a été atteinte en gérant au départ moins de $10M d'investissements[1366].

La tendance dans ce secteur est de faire de plus en plus appel à des sources de données dites alternatives, à savoir extérieures aux entreprises analysées. L'idée est d'obtenir des indicateurs périphériques et croiser des données de sources variées. Elles peuvent par exemple provenir des réseaux sociaux, d'indicateurs économiques divers, de la météo ou d'observations satellite, de données de trafic Internet, de suivi de prix de matières premières ou produits finis, etc. Là-dessus, le machine learning fait son travail pour trouver des corrélations entre toutes ces données, la santé future des entrepries et, ultimement, de leur valeur en bourse[1367].

D'autres solutions à base d'IA sont tout aussi intrigantes. C'est le cas d'un chercheur japonais, Kyoshi Ozumi, de l'université de Tokyo, qui analyse les expressions des visages des gouverneurs des banques centrales pour interpréter leurs déclarations et vérifier leur sincérité. Bref, une sorte de détecteur de mensonges visuel. D'autres analysent les déclarations écrites avec des outils d'analyse du langage. Tout cela semble encore très artisanal[1368].

Accern (USA, $19M) propose une solution de création de workflows pour la finance sans développement logiciel. Elle consiste surtout à analyser des flots de documents publics pour faire de la recherche ciblée d'information sur les entreprises dans lesquels les fonds sont investis. La solution exploite donc surtout des briques de traitement du langage, avec une interface permettant de paramétrer ses besoins[1369]. C'est moins magique mais pas survendu.



Des solutions sont même censées permettre d'identifier des startups dans lesquelles investir. Je les aborde avec une prudence extrême car les facteurs clés de succès des startups sont avant tout humains et mal encodés dans ce que l'on peut trouver en scrappant Internet.

[1366] Voir Can an artificial intelligence learn to beat the stock market? par William D.Cohan dans Fast Company, mai 2020. Description du procédé : "*The software he runs is a type of theorem prover, a nondeterministic algorithm that can look at a data set and generate a hypothesis to interpret what it sees. Similar to how the human brain breaks information into chunks in order to form heuristics about the world*".

[1367] Voir AI Is Trading's Next Key Differentiator par Hazem Dawani, décembre 2019 ainsi que L'IA est en train de révolutionner la gestion d'actifs des banques françaises par Antoine Crochet-Damais, octobre 2020, qui évoque de nombreux cas d'applications dans des banques françaises.

[1368] Voir Prédire les politiques monétaires grâce à l'IA, un graal pour la finance par Antoine Hasday, 2020.

[1369] Voir Artificial Intelligence Startup Accern Raises $13 Million In Series A To Help Enterprises Adopt AI More Easily par Igor Bosilkovski, mai 2020.

Le fonds d'investissement de Hong-Kong **Deep Knowledge Venture** spécialisé dans la santé communiquait avec fracas en 2014 sur le fait que l'un des membres de son comité d'investissement était une IA. C'est en fait une solution extrêmement pointue, développé par Aging Analytics (UK), spécialisé dans la recherche sur les technologies de lutte contre le vieillissement.

Mais indiquer qu'une IA a une place dans un board est aussi stupide que si un cabinet d'expert comptable indiquait qu'un PC équipé d'un tableur Excel était un de ses employés. Aussi sophistiquée soit-elle, une solution d'IA reste un outil d'aide à la décision, comme dans n'importe quel autre processus de décision qui s'appuie sur la rationnalité de données. Il faut toujours y ajouter un peu d'intuition et de connaissance humaine !

Le russe **TalentBoard** et sa solution Test4startups (ou T4S) aide de son côté les startups à valider leur business plan et leur marché. Le système analyse le marché visé, les concurrents, la liste des investisseurs potentiels et produit une évaluation du projet. Il faut prendre cela avec du recul car cela ne suffit évidemment pas à prendre une décision éclairée d'investissement dans une startup, surtout d'innovation « de rupture ».

Le fonds suédois **EQT Venture** gère un fonds d'investissement dans les startups de 566M€. Il fait partie de la société d'investissement EQT qui gère $38B d'actifs. EQT Ventures exploite sa propre solution d'IA dénommée Motherbrain pour faire ses choix d'investissements et surtout, détecter les startups les plus prometteuses qui sont « sous le radar » et, ce qui est le plus important, avant les autres fonds[1370]. L'IA exploite une myriade d'informations comme les financements de startups, les classements sur Internet et l'activité dans les réseaux sociaux plus des données générées par ses propres analystes.

Le machine learning exploité est non supervisé pour identifier des tendances et supervisé pour exploiter des données déjà labellisées. Motherbrain est aussi exploité par le fonds pour accompagner les startups dans lesquelles il investi[1371].

Ayomi (2017, France, 1M€) est une plateforme de financement en prêt et en capital qui indique s'appuyer sur une IA pour trouver les bons investisseurs. Comme toute IA, elle a besoin de données pour ce faire : votre carnet d'adresses dans lequel elle va fouiller pour trouver les investisseurs probablement avec une solution de machine learning multiparamètres. IA ou pas, cela ne fonctionne évidemment pas à tous les coups.

[1370] Je suis évidemment très réservé sur ce genre de méthode pour trouver les véritables pépites. Les données génèrent ce que l'on appelle le biais du rétroviseur : elles ne permettent pas de voir en avant ! Qui plus est, les facteurs clés de succès d'une startup sont des signaux faibles difficiles à détecter. Les données permettent cependant sans doute d'éliminer les plus mauvais projets, même si un coup d'œil sur le dossier et sur l'équipe permet aussi de le faire. C'est la détection des meilleurs projets qui est difficile, surtout lorsque l'on sait que le succès des meilleurs dépend aussi de la chance et de phénomènes sociétaux. Même si ces derniers peuvent être détectés dans les réseaux sociaux.

[1371] Les fonds américains **SignalFire** et anglais **InReach Ventures** utilisent aussi du machine learning pour choisir les startups dans lesquelles ils investissent.

Mattermark (2012, USA, $17,2M) est un fournisseur de base de données d'entreprises servant à la détection de prospects pour la vente en tout genre et pas seulement pour les investissements par les sociétés de capital risque. Là encore, il faut regarder de près les données qui alimentent leur système et prendre l'ensemble avec des pincettes.

Clearbanc (2015, Canada) propose un nouveau mode de financement non-dilutif des startups qui font du ecommerce ou du cloud. Ce financement allant de $10K à $10M est censé servir à faire de l'acquisition de clients via les grandes plateformes en ligne. Il est fourni à un taux d'intérêt fixe de 6% et de 12% si le financement sert à autre chose qu'à de l'acquisition client. Clearbanc se finance aussi par du revenue sharing. La sauce magique de l'histoire est une IA à base de machine learning qui permet à Clearbanc d'évaluer l'état de l'activité et la trajectoire de la startup via les plateformes ecommerce, de référencement et de paiement en ligne telles que Stripe, Shopify et Google Adwords. Leur modèle de la *term-sheet* générée en 20 minutes permet de rationnaliser la prise de décision d'investissement et, notamment, de mieux financer les startups créées par des femmes qui sont parfois défavorisées par les investisseurs classiques[1372].

Il faut aussi compter avec **Kensho Technologies Inc** (USA) qui s'appuie sur l'IA en général pour fournir des outils d'analyse aux investisseurs dans les banques.

Et ne j'ai pas cité les projets et startups qui cherchent à prédire les cours de la bourse ou d'indices boursiers à base d'IA, comme **QuantCube**[1373] (2013, France, $5,1M)[1374] ! Nous avons aussi **Cognivi Labs** (2016, USA, $3,1M) qui utilise force machine learning à base de calcul de la confiance des consommateurs pour prédire les évolutions des cours de bourse de grandes entreprises, avec une étude de cas sur Facebook et une autre sur Shechers.

C'est à prendre avec des pincettes car ils ne publient évidemment pas les cas où leur prévision n'a pas bien fonctionné ! Bref, une belle application du biais du survivant.

Les fonds d'investissement de type hedge funds qui exploitent du machine learning sont en tout cas plus performants que les fonds traditionnels.

[1372] Voir Clearbanc Is Using AI To Level The Playing Field In Funding, par Sandra Ponce de Leon, avril 2019.

[1373] Quantcube fait des prévisions d'indices économiques pour les hedge funds. Ils font du croisement de données météo, satellites permettant d'identifier la construction de villes et le suivi des stocks de pétroles comme Orbital Insights, le suivi de l'activité hôtelière, le suivi des navires de commerce, le tout permettant de déterminer la croissance économique d'une région ou d'un pays.

[1374] Voir aussi Forecasting the Stock Market Index Using Artificial Intelligence Techniques de Lufuno Ronald Marwala (166 pages).

C'est ce qui ressort de ce chart de comparaison de 13 fonds utilisant du ML (en bleu en haut) avec des fonds classiques (les autres courbes), issu d'**Eurekaedge** (2001, Singapour)[1375]. On observe cependant que le ML a permis de créer une forte différence alors qu'il était encore rarement utilisé, entre 2010 et 2016 et que, depuis 2018, la performance de ces fonds s'est tassée. A savoir qu'un investissement réalisé dans les hedge funds à base d'IA début 2017 n'a visiblement généré aucun avantage. Reste à l'interpréter ce qui n'est pas évident ! Il est possible que des méthodes de ML soient aussi utilisées par les autres hedge funds.

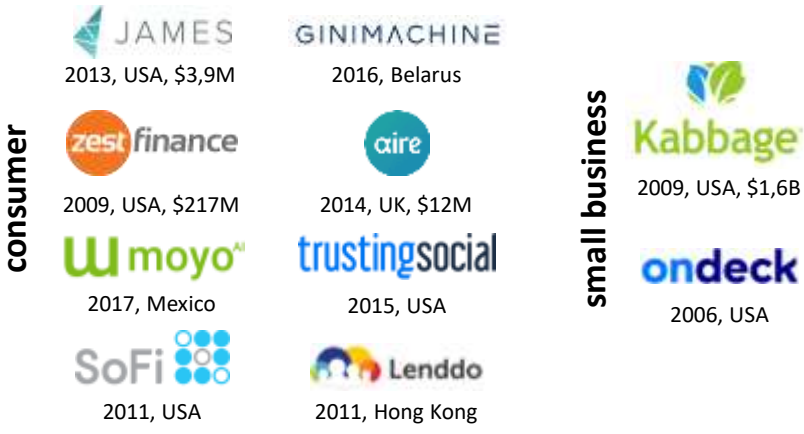
Dans le même registre, une étude de chercheurs suisses et suédois de 2020 comparait la performance de trois classes d'investisseurs : des business angels novice, des business angels expérimenté et une solution à base de machine learning. Le résultat montre que le machine learning est meilleur que les novices mais moins bon que les investisseurs expérimentés. Ce n'est pas une grande surprise. Au moins, on se rend compte que le machine learning peut améliorer la performance des investisseurs novices. A ceci près que si tous les investisseurs se mettent à utiliser les mêmes outils, le marché se nivelle par le haut et la différence humaine reprend le dessus[1376].

Gestion des risques

La gestion des risques porte sur le *credit rating*[1377] d'emprunteurs basé sur les informations disponibles sur les réseaux sociaux avec **TrustingSocial** (2015, USA), **Lenddo** (2011, Hong Kong, $14M) ou **Kreditech** (2012, Allemagne, $497M), ce qui n'est pas sans poser diverses questions sur le respect de la vie privée.

De nombreux services de crowlending tels que **Kabbage** (2009, USA, $1,6B) qui cible le marché des PME aux USA et garantit des prêts qui font en moyenne $200K, **OnDeck Capital** (2006,USA) qui cible aussi les PME, **SoFi** (2011, USA) qui gère des prêts pour étudiants, **Lending Club** (2007, USA) et **ZestFinance** (2009, USA, $217M) font aussi appel au machine learning pour le credit rating, ce dernier en se focalisant sur les emprunteurs qui ont une faible empreinte en ligne. **OakNorth** (2015, UK, $448,5M) fait des prêts aux PME dont le risque est aussi évalué avec de l'IA avec leur solution ACORN.

Le *credit rating* à base d'IA est aussi proposé par **James** (2013, USA, $3,9M), **Aire** (2014, UK, $12M) qui se focalise sur les jeunes emprunteurs sans 'casier' financier. **GiniMachine** (2016, Belarus) a une offre voisine. **Upstart** (USA, $85,7M) propose des prêts aux particuliers. Le machine learning permet de générer les taux d'intérêts et autres conditions de prêts automatiquement.



[1375] Eurekahedge analyse la performance de plus de 30 000 fonds d'investissement dans le monde. Voici le lien permettant de reconstituer le chart : source.

[1376] Voir Do Algorithms Make Better — and Fairer — Investments Than Angel Investors? par Torben Antretter et al, novembre 2020.

[1377] Voir Application of Artificial Intelligence Techniques for Credit Risk Evaluation de Ahmad Ghodselahi et Ashkan Amirmadhi qui décrit une méthode de credit rating à base d'arbres de décision, de machine learning et deep learning exploitant une dizaine d'agents différents. Voir aussi Credit Scoring Models by AI Companies: A Comprehensive Guide, Credit Scoring Using Machine Learning, 2013 (381 pages) et Artificial intelligence and machine learning in financial services, 2017 (45 pages).

Smart Finance (2013, Chine) demande aux utilisateurs d'accéder à leurs données mobiles. Le machine learning vient à la rescousse pour identifier des corrélations entre des centaines de paramètres de la vie numérique mobile des mobinautes et leur fiabilité d'emprunteur. Cela va jusqu'à récupérer le niveau de charge de la batterie. Une faible charge chronique serait un paramètre corrélé avec une propension à ne pas rembourser ses prêts ! Voilà de quoi créer de beaux biais des données !

LendUp (2011, USA, $361M) propose des prêts aux particuliers en optimisant leurs taux d'intérêts de prêts. C'est une forme de système de micro-sub-primes, attribuant des crédits à des particuliers qui ne peuvent pas accéder aux crédits dans les circuits traditionnels. Comme c'est une activité plus risquée que les crédits traditionnels, la startup doit se couvrir avec des algorithmes qui limitent les risques en question !

L'IA détecte les clients potentiels qui ont des comportements « sains » comme…le fait de bien rembourser ses crédits et d'avoir un budget équilibré ! On peut se demander pourquoi il faut de l'IA pour déterminer cela !

En mode b2b, il est aussi réalisé par **JPMorgan Chase** depuis 2017 avec sa solution COiN, qui utilise le machine learning pour analyser leurs 12 000 contrats de crédit commercial annuel en quelques secondes, faisant sur le papier gagner 360 000 heures par an à la banque d'investissement.

RCI DIAC, filiale du groupe **Renault** s'appuie sur **Score4Biz** (2012, France) pour déterminer en temps réel un score d'octroi de crédit. Elle a déployé une solution de machine learning basée sur des travaux de recherche issus de Telecom ParisTech et de l'ENS Paris Saclay.

Les engagements de services comprennent un temps de réponse de 250 ms en moyenne avec un engagement contractuel de moins d'une seconde.

Transax (2006, USA) est une filiale de Pineapple Payments qui gère des systèmes de paiement et propose des outils de scoring qui exploitent comme il se doit du machine learning.

Capnovum (2014, UK) propose une solution logicielle qui cartographie dynamiquement les endroits dans le monde où la réglementation financière évolue (vidéo). Ils suivent les tendances et les effets de suivisme entre régulateurs de régions et pays. Une IA non précisée est utilisée pour identifier des relations entre ces évolutions de régulations et les activités économiques de l'entreprise ou l'état de ses investissements. Elle doit probablement au minimum exploiter des briques de traitement du langage pour analyser les textes. Cela aide les entreprises internationales à bien gérer leur respect des réglementations. L'histoire ne dit pas si cela permet au passage de faire de l'optimisation fiscale respectueuse des réglements !

Fortia (2012, France) utilise dans sa solution Innova des algorithmes qui comprennent les réglementations et vérifient leur application en calculant notamment les ratios de couverture de risques d'investissements financiers.

Leur solution exploite différentes briques d'IA : du traitement du langage et de la vision artificielle pour analyser les textes et des régressions non linéaires sur les données pour évaluer des risques (vidéo).

Enfin, l'évaluation du risque doit être aussi faite côté client en b2b. C'est la mission de **Neuroprofiler** (2016, France) qui propose un jeu en ligne qui permet de qualifier le profil d'investisseur financier de clients, pour respecter la règlementation financière européenne MiFIDII. Le tout s'appuie sur du deep learning qui associe les réponses au jeu à un profil type financier en terme de capacité et de compréhension de prise de risque.

Détection de fraudes

La détection de fraude est un cas d'application classique du machine learning. Les fraudes sont détectées en collectant un maximum d'information sur les payeurs et en identifiant les « patterns » de mauvais payeurs.

C'est ce que propose **Sift Science** (2011, USA, $106,6M) avec une offre de sécurisation généraliste destinée aux banques et commercants et aussi **Riskified** (2012, Israël, $63,7M), qui est focalisé sur les sites marchands.

Visa exploite des réseaux de neurones depuis 1993 pour détecter en temps réel le risque des transactions. C'est pour cela que vos débits sont parfois refusés lorsque vous êtes en déplacement dans des zones inhabituelles. Le système s'est amélioré avec le temps pour éviter les faux positifs et éviter plusieurs dizaines de milliards d'Euros de fraudes au paiement représentant 0,1% des transactions. Ils exploitent différentes données anonymisées comme les habitudes et lieux d'achats avec plusieurs centaines de caractéristiques. Ce qui va sortir de vos habitudes sera considéré comme suspect, avec les limites du genre si vous en changez brutalement ! Les scores de risques des transactions sont communiqués aux banques en une milliseconde modulo la latence des moyens de télécommunication utilisés. Ce sont les banques qui décident en dernier ressort d'accepter ou pas vos transactions.

Les banques doivent aussi passer au peigne fin toutes les transactions de plus de $10K pour détecter le blanchiment d'argent sale. Là encore, il faut faire appel à du machine learning voire du deep learning pour trier les centaines de milliers de transactions.

C'est ce que propose de faire la startup **Simularity** (2011, USA) qui aide à détecter les anomalies dans de nombreux marchés verticaux, dont la finance[1378]. Cela doit probablement faire appel à un mix de machine learning supervisé (pour détecter des anomalies connues) et non supervisé (pour les anomalies nouvelles et non labellisées).

Quantexa (2016, UK, $23,3M) aide à lutter contre la criminalité financière, le blanchiment d'argent sale et à réduire les risques de crédit. Shell Oil fait partie de leurs premiers clients (exemple de cartographie *ci-contre*).

DataVisor (2013, USA, $54,5M) fait de la détection d'anomalies dans les transactions financières à base de machine learning.

Biocatch (2011, Israël, $41,6M) détecte le vol d'identité dans les transactions financières en ligne à l'aide de biométrie comportementale. Elle s'entraîne en analysant le comportement des utilisateurs lors de leurs interactions avec les outils numériques pour identifier leur usage par d'autres personnes ou par des bots (vidéo).

La banque **Santander** utilise la reconnaissance vocale pour sécuriser les transactions, avec la solution de la startup **Fonetic** (2006, Espagne) qui analyse les émotions dans les conversations téléphoniques.

Deux startups spécialisées dans les applications de gestion de la conformité des transactions, l'Américaine **Lucid** (2004, USA, $15M, acquise par Acuity Brands en février 2018) et **Feedzai** (2009, USA/Portugal, $76M) utilisent toutes deux le machine learning pour détecter 80% des fraudes.

Chekk (Hong-Kong) est une startup crée par des Français qui fournit une solution de bout en bout pour vérifier l'identité de clients dans le secteur de la finance. Ils exploitent un mix de signatures électroniques et d'outils de vérification de l'identité s'appuyant sur du traitement de l'image pour la détection de faux documents et des pièces et photos d'identité des clients. Ces outils sont ensuite exploités dans la relation avec les clients.

[1378] Voir leur intéressant livre blanc Artificial Intelligence (AI) for Financial Services.

Coinbase (2010, USA) gère un portefeuille électronique de Bitcoin et s'appuie sur de l'IA pour éviter la fraude et l'usurpation d'identité[1379]. Cela ne doit pas être bien facile à évaluer !

Chatbots

Comme dans le commerce en ligne, les chatbots sont très à la mode et donnent lieu à une pléthore d'offres de startups et de banques en ligne.

Le marché des chatbots financiers se structure en gros en trois types d'acteurs : les startups proposant leur chatbot grand public qui s'interface plus ou moins avec les services bancaires de l'utilisateur, les startups qui proposent des chatbots en marque blanche pour les banques et les banques qui proposent un chatbot à leurs propre clients.

Les startups de la première catégorie sont très nombreuses avec **Cleo** (2015, UK, $40,9M), **Chip** (2016, UK), **Digit** (2013, USA, $63,8M), **Trim** (2015, USA, $2,2M), **Penny** (2015, USA, $1,2M, acquis par Credit Karma en 2018) et **Dyme.co** (2014, USA, 650K€). Il va sans dire qu'il y aura de la casse dans ce secteur, comme dans n'importe quel marché dans lequel s'engouffrent des dizaines de startups faiblement différentiées, et souvent, pas très bien financées. Il est probable que celles qui s'en sortiront le mieux seront celles dont l'écosystème sera le plus dense, avec les interfaces avec les banques, d'autres services financiers voire commercants.

Côté marque blanche, nous avons avec notamment l'Américain **Kasisto** (2013, USA, $43,5M)[1380] et son chatbot qui est notamment utilisé par MasterCard, **Finn.ai** (2014, Canada, $13,9M) et **Personetics Technologies** (2010, Israël/USA, $18M) qui couple un agent conversationnel avec du prédictif sur les besoins des clients.

Et puis, nous avons bien évidemment **IBM Watson** qui peut être utilisé pour créer des chatbots, avec l'accompagnement des équipes services d'IBM. Il est notamment mis en œuvre en France par **Crédit Mutuel CIC**, non sans quelques réactions négatives des syndicats de salariés inquiet pour l'impact sur l'emploi dans les agences. Le chatbot mis en œuvre dans cette banque est un outil pour 20 000 salariés[1381] qui trie 350 000 mails entrants par jour et divers chatbots permettant d'interroger les bases de connaissance de la banque sur ses différents métiers (épargne, assurance auto, assurance santé et prévoyance). L'ensemble est développé par l'équipe de la Cognitive Factory avec 75 personnes dont une moitié provient d'IBM. Le projet aurait coûté 40M€ étalés sur 5 ans.

Aux USA, des chatbots ont été lancés par **Bank of America** avec Erica qui est plutôt structuré comme un système de recommandation et est aussi commandable par la voix (*ci-dessous à droite*), et **American Express** sur Facebook Messenger depuis fin 2016 et **Wells Fargo** depuis mi 2017.

Au Royaume Uni, **Barclays** a lancé son Launchpad qui permet d'exécuter des tâches de son application mobile classique en mode dialogue, sans que cela soit d'ailleurs plus efficace. Le chatbot de la **Royal Bank of Scotland** est développé avec IBM Watson. On trouve aussi des chatbots chez **Santander** en Espagne[1382] et **Swedbank** en Suède avec son agent conversationnel textuel et vocal développé par l'Américain Nuance.

[1379] Voir AI at scale at Coinbase, septembre 2018.

[1380] Avec plus ou moins de bonheur, voir : https://www.wired.com/2016/06/new-banking-ai-now-chatbots/.

[1381] Confirmé par une décision de la Cour de Cassation du 12 avril. Voir Cour de cassation, civile, ch. sociale, arrêt du 12 avril 2018 qui a considéré que l'impact du déploiement d'IBM Watson sur les conditions de travail des salariés étaient mineures. Voir aussi cette intéressante étude de l'impact de l'IA sur l'emploi dans la finance Impact of Artificial Intelligence (AI) on Financial Job Market – An Introduction 2018 (44 pages). Le chatbot permettrait de faire gagner 10 minutes par jour aux collaborateurs de la banque de détail.

[1382] Voici la source de nombre de ces différents exemples : « Artificial Intelligence in Digital Bankin » de MAPA, novembre 2016.

En Chine, les principales banques comme la **Bank of China** ont un chabot intégré dans **Wechat**. Enfin, la banque **OCBC** de Singapour a créé un chatbot dénommé Emma (tous les prénoms féminins y passent…) spécialisé dans l'attribution de prêts pour la rénovation de logements[1383].

En France, **Orange** a lancé en 2017 son offre bancaire Orange Bank, qui s'appuie fortement sur une application mobile et l'usage intensif d'IA, notamment dans un chatbot. Orange Bank a été développé, entre autres, autour d'IBM Watson. Le lancement en novembre 2017 a été chaotique et la montée en charge commerciale du service est sujette à caution. L'activité était plus déficitaire que prévu en 2019.

Mais au juste, est-ce que les chatbots fonctionnent bien et sont appréciés des utilisateurs ? Rien n'est moins sûr ! La qualité d'un chatbot dépend surtout des processus qui ont été intégrés dans sa base de connaissances. Ils sont souvent très limités et les capacités de dialogue des chatbot ne vont pas très loin.

Un bon chatbot doit laisser la main à un véritable interlocuteur lorsqu'il détecte que la communication ne se déroule pas convenablement et les banques n'ont pas encore mis en place de genre de solution. Le message marketing mis en avant est toujours ambigu : les chatbot permettent d'améliorer la satisfaction client. Ils servent surtout à réduire les coûts de la banque de détail[1384].

Assurance

Comme la finance, le métier d'assureur tourne beaucoup autour de la donnée. L'optimisation de la gestion du risque est encore plus critique que dans les services financiers puisqu'elle fait partie du cœur de métier.

distribution

optimisation des **canaux de distribution** (ML)

optimisation de l'approche de **nouveaux marchés** (ML)

recommandations de **packages d'assurances optimisés** (ML)

création d'offres de **cross-selling** (ML)

gestion des risques

évaluation personnalisée des risques et des offres (ML)

ciblage des clients à risques faibles (ML)

stratégies de réduction ciblée des risques (modes de vie pour la santé, style de conduite pour l'automobile, protection de l'habitat pour la sécurité) (ML, DL)

gestion des sinistres

qualification des sinistres via analyse de photos et des textes de description (DL)

détection des fraudes (ML)

comparaison et **optimisation des devis** de fournisseurs agréés (ML)

relation client accélérée via des chatbots (NLP, DL)

aide à la **saisie de déclaration** de sinistres (ML)

prévisions des vagues de sinistres pour optimisation d'allocation du capital (ML)

Dans l'assurance, l'IA intervient dans tout le cycle produit : pour segmenter ses clients, créer des produits, cibler ceux qui présentent le moins de risques, proposer les bonnes offres aux clients, gérer des actions préventives de réduction des risques chez les clients, gérer les expertises et détecter les fraudes[1385]. La relation client fait, comme dans les banques, aussi appel aux chatbots[1386].

[1383] Source : OCBC bank launches first artificial intelligence powerd home & renovation loan specialist, avril 2017.

[1384] Voir Bots Aren't Ready To Be Bankers de Forrester, août 2016.

[1385] Voir Emerging Technologies Transforming the $4tn Insurance Industry, de CommerzVentures, 2016. Insurance 2030—The impact of AI on the future of insurance de Ramnath Balasubramanian, Ari Libarikian et Doug McElhaney, avril 2018 et AI in insurance Pega (13 pages).

[1386] Voir Why Chatbots Are Taking over the Insurance Industry, août 2017.

En France, les grandes assurances se sont lancées dans l'adoption de l'IA, que ce soit par exemple chez **Axa**, **Generali**[1387] ou **Harmonie Mutuelle**[1388]. De nombreuses startups se sont évidemment aussi lancées dans ce secteur et en exploitant des briques d'IA.

Distribution

Commençons par la distribution de contrats d'assurance.

Zelros (2016, France, 4,5M€) propose un chatbot spécialisé dans l'assurance pour la relation client en avant-vente et pour la vente de contrats et la gestion des déclarations de sinistres. L'assistant personnel est exploité par les conseillers commerciaux. Il est alimenté par des outils prédictifs à base de machine learning. La solution est déployée chez MAIF, Natixis Assurances, Groupama, LBP et Matmut.

Gan Prévoyance, filiale du groupe **Groupama** s'appuie sur **DreamQuark** (2014, France, $3,5M) pour déployer une solution d'IA de deep learning afin d'améliorer la connaissance de ses clients et optimiser son développement commercial[1389].

Minalea (2015, France) a développé un chatbot de vente de contrats d'assurance, l'Assistant Commercial Intelligent pour les conseillers des courtiers en assurance. Il fait l'inventaire des garanties et services des produits d'assurance du marché et permet de trouver la bonne offre rapidement avec l'argumentaire de vente associé. Ce système est proposé pour l'assurance automobile, les multi-risques habitations, la couverture d'emprunt, l'assurance incendie.

Amplyfi (2015, UK, £800K) collecte plein de données ouvertes pour aider les assurances à « mieux connaître » leurs clients et définir des polices d'assurance personnalisées avec leur solution Data-Voyant[1390].

Riskgenius (2015, USA, $3M) est à l'origine d'une solution de traitement du langage qui gère le cycle de création et modification de contrats d'assurances. C'est une solution utile pour les agents d'assurance qui peuvent ainsi plus facilement comparer les polices d'assurances distribuées.

IBM Watson a été mis en œuvre dans **Insurance Assistant** de l'USAA (United Services Automobile Association), un agent conversationnel qui permet aux clients de cette assurance dédiée au personnel militaire US de s'y retrouver dans ses offres et services.

Une startup peut ambitionner de remplacer les compagnies d'assurance traditionnelles. C'est le cas de **Lemonade** (2015, USA, $480M), est une société d'assurance en ligne pour propriétaires et locataires basée à New York qui s'appuie fortement sur l'IA dans tous ses processus et se passer d'intermédiaires (les courtiers), y compris des chatbots dans la relation client (vidéo). La société utilise les sciences comportementales pour limiter la fraude. Par exemple, les constats sont réalisés par vidéo en ligne qui sont ensuite exploitées avec des solutions d'IA de détection d'émotion du style de celle d'**Affectiva** (2009, USA, $60,3M).

[1387] Voir Generali fait de l'IA un levier de son plan de transformation, par Xavier Biseul dans Zdnet.fr, juillet 2019. Au menu, le chatbot client Leo, la gestion des flux entrants d'emails avec une solution construite autour des briques de traitement du langage IBM Watson, de la RPA pour l'automatisation des processus internes et du machine learning pour la réduction des risques de fraudes. Mais tout cela en maintenant les collaborateurs dans la boucle.

[1388] Voir Comment Harmonie Mutuelle met en place sa stratégie IA par Didier Barathon, 2018. Avec de la détection de fraude, l'extraction de données dans les documents, la RP, un chatbot interne, la reconnaissance vocale et la visualisation des données.

[1389] DreamQuark développe des solutions d'intelligence artificielle à base de réseaux de neurones et de deep-learning avec des mécanismes d'auto-apprentissage capables d'explorer tous seuls tous types de données de les traiter. La startup propose des outils d'analyse via sa plateforme Brain qui permet d'explorer, optimiser et valoriser les données structurées (bases de données) et non-structurées (images, sons, voix) dans les secteurs de la finance et de l'assurance. Ils ont notamment comme client BNP Parisbas Wealth Management.

[1390] Voir How Your Insurance Quote Is Powered By Artificial Intelligence par Gina Clarke dans Forbes, janvier 2019.

Natixis Assurances utilise enfin un système de RPA (robotic process automation) pour automatiser la cloture des contrats et la communication par email avec ses assurés.

Gestion des risques

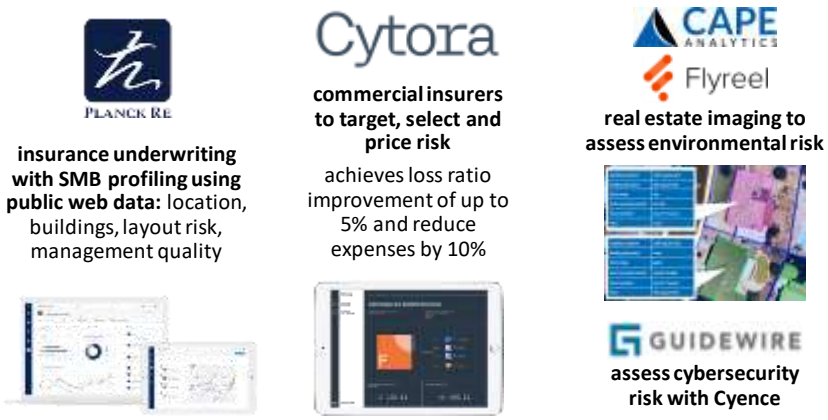
Du côté de la gestion des risques, voici quelques offres et études de cas[1391]. Elles portent sur l'évaluation des risques économiques, de cybersécurité et environnementaux pour l'immobilier.

Planck Re (Israël, $12M) fait de l'analyse de risques assuranciel exploitant des sources ouvertes sur les entreprises (images, textes, vidéos, sentiments, réseaux sociaux).

Cytora (2014, UK, $41,4M) réalise une analyse de risque et va jusqu'à aider à la tarification des assurances pour minimiser les pertes. Ils fournissent une partie de leurs services sous forme d'APIs destinées aux applications d'actuariat. Leur solution est exploitée dans une filiale d'Axa. Cette start-up est une spin-off de l'Université de Cambridge.

Guidewire Software (2001, USA, $24,8M) a fait l'acquisition de **Cyence** (2014, USA, $40M) en octobre 2017. C'est une solution à base d'IA qui évalue les risques en matière de cybersécurité pour les entreprises clientes des assurances en exploitant de grandes masses de données ouvertes ou propriétaires tierces-parties. Cela permet aux assureurs de mieux adapter leurs offres aux clients entreprises et à minimiser les risques associés.

Cape Analytics (2014, USA, $31M) analyse les bâtiments par de l'imagerie aérienne permettant de qualifier la géométrie des bâtiments, la qualité des couvertures et celle de la végétation. Cela permet d'évaluer les risques dans l'immobilier. **Flyreel** (2016, USA) propose une offre voisine.



Côté assureurs, **Allianz**, exploite les données internes et externes à l'entreprise pour identifier la situation du client ou prospect. Il peut par exemple détecter que le client gare habituellement son véhicule dans une zone d'une ville où les vols sont plus nombreux que la moyenne et proposer une assurance contre le vol. Le tout est exploité dans un chatbot qui s'appuie sur IBM Watson.

Les assurances sont partenaires de fournisseurs de solutions de maison intelligente pour réduire les risques de sinistres dans les logements ainsi qu'avec divers fournisseurs de solutions de santé pour réduire les coûts de santé, pour ce qui est des assureurs santé, surtout aux USA.

Dans l'assurance auto, on peut aussi encourager les conducteurs à faire auditer leur mode de conduite via des capteurs au standard CAN-2 dont l'offre est très abondante, notamment avec **Oocar** (2015, France).

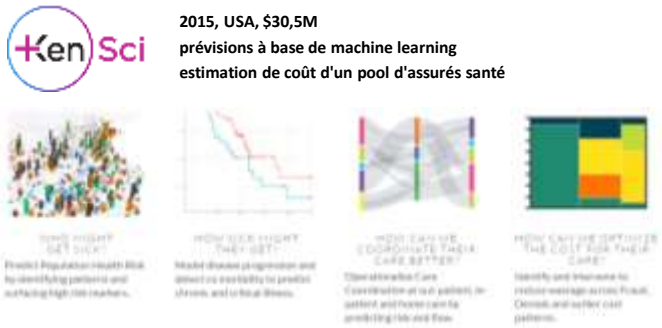
[1391] Sur la gestion des risques dans les assurances, voir AI and risk management Innovating with confidence de Deloitte (32 pages), Impact of Artificial Intelligence on Reinsurance Sector Scor 2018 (36 pages) et The Rise Artificial Intelligence: Future Outlook and Emerging Risks, 2017 (24 pages).

Gestion des sinistres

La gestion des sinistres bénéficie de nombreuses briques de l'IA : le routage automatique des emails des assurés, l'usage de la vision artificielle pour évaluer les sinistres dans les logements ou l'automobile, les chatbots et autres agents vocaux et le machine learning pour détecter différentes formes de fraudes[1392].

Element.ai (2016, Canada, $250M) est un éditeur de logiciels d'applications métiers qui s'appuient sur l'IA. L'un des marchés qu'ils gèrent est la gestion des sinistres dans l'assurance, pour la couverture des risques sur les biens physiques (immobilier, automobile). Ils gèrent la collecte d'informations structurées et non structurées sur les contrats, les clauses de couverture et les sinistres pour aider les assureurs à prendre position rapidement sur chaque dossier. La startup a été acquise en novembre 2020 par l'opérateur de services en cloud **ServiceNow** (2004, USA, $87,3M) et pour $230M. Il semblerait que leur difficulté à créer de véritables produits logiciels et une approche de vente scalable en soit à l'origine[1393].

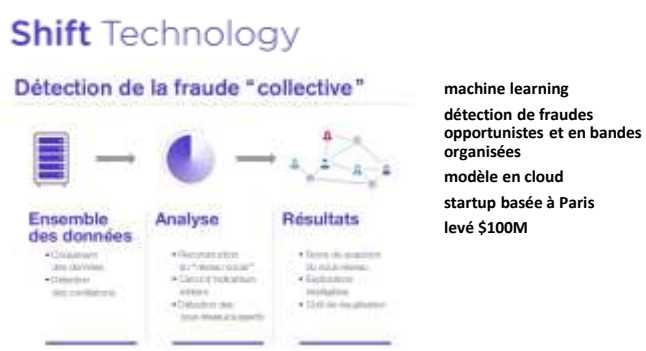
KenSci (2015, USA, $30,5M) est spécialisé dans les prévisions à base de machine learning. Il permet notamment aux assureurs santé de prévoir la structure de coût d'un parc d'assuré et de lancer des programmes de soin les diminuant (vidéo). Leurs modèles de machine learning permettent au passage de prédire l'espérance de vie de leurs clients et d'ajuster les « *dispositions à prendre pour optimiser leur fin de vie* ». Réjouissant[1394] !



Understory (2012, USA, $9,5M) est une startup dans les objets connectés qui fournit des capteurs d'environnement (humidité, température, vent, précipitations) assimilables à des stations météo miniatures qui permettent d'auditer a posteriori l'origine de dégâts d'origine météorologiques.

Tensorflight (2016, USA) utilise l'IA et l'imagerie aérienne pour automatiser l'inspection aérienne de biens immobiliers et évaluer d'éventuels destructions liées à des catastrophes naturelles[1395].

Shift Technology (2014, France, $100M) est spécialisé dans la détection de fraude organisée avec une solution en cloud. Elle permet par exemple d'identifier des redondances dans les scénarios et éléments fournis par les assurés, comme des photos qui seraient toujours les mêmes pour décrire les sinistres ! Ils sont déjà plus de 45 clients dont AG2R La Mondiale et démarrent leurs opérations à New York.



[1392] Voir AI in Insurance: Business Process Automation Brings Digital Insurer Performance to a New Level, juillet 2018, qui porte sur la gestion des remboursements.

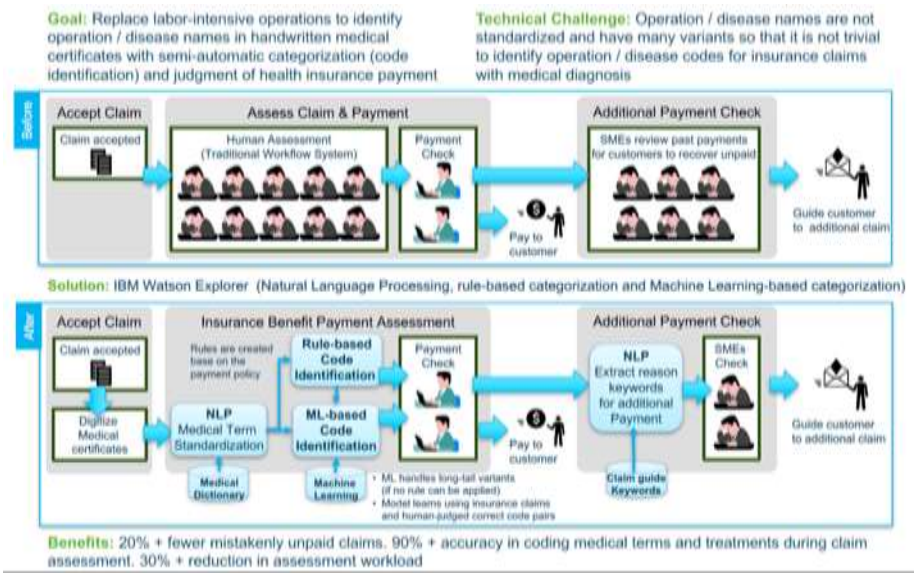
[1393] Voir ElementAI, un « fleuron » de l'intelligence artificielle ? par Yves Gingras et Maxime Colleret, décembre 2020.

[1394] Voir KenSci research paper on End-of-Life Prediction to be presented at the AAAI 2018 conference, février 2018.

[1395] Voir Roof Damage Assessment using Deep Learning, 2018 (6 pages) qui décrit les méthodes d'analyse de toits à base de deep learning, utiles pour les assurances en cas de catastrophe naturelle.

La détection de fraude à base d'IA est utilisée par l'assurance **Manulife** au Canada avec une solution développée en partenariat avec l'Université de Waterloo ainsi que par **Reliance Nippon Life Insurance** en Inde. Ce dernier évite ainsi de couvrir avec des assurances décès des personnes dont la mort prochaine est programmée à moins de trois ans !

En 2017, l'assurance japonaise **Fukoku Mutual Life Insurance Co** faisait parler d'elle en annonçant prévoir de réduire de près de 30 % le personnel de son équipe d'évaluation des paiements grâce à une IA basée sur les briques logicielles d'IBM Watson et lancée en janvier 2017[1396]. Soient 34 personnes (qui étaient en CDD de 5 ans) sur 131 (ce qui fait 26%...).



Cette IA exploite les divers documents médicaux pour gérer les remboursements de ses assurés santé en conformité avec le niveau de couverture. Plus d'un an plus tard, il est difficile de savoir si la solution a tenu ses promesses. Il faut être toujours prudent avec ce genre d'annonce !

Toujours en 2017, un autre assureur japonais, **Nippon Life Insurance Co** mettait en production une AI d'analyse des meilleurs plans de couverture santé pour les clients grand public, exploitant les données de 40 millions de contrats.

Les assurances font aussi appel à la reconnaissance d'image dans le cadre d'expertises, notamment automobiles ainsi que pour scanner les constats.

Tractable (2014, UK, $34,9M) propose une solution d'inspection visuelle d'automobiles basée du deep learning pour de la classification automatique (vidéo).

Tchek (2016, France) a développé un scanner à 360° de véhicules qui permet d'en faire un diagnostic complet, notamment pour identifier tous les défauts de la carrosserie. Les cas d'usages les plus évidents sont pour les assureurs mais aussi les loueurs de véhicules. Les véhicules sont inspectés en 5 secondes.

Fotonower (2014, France) propose une solution de diagnostic à distance d'accidents de véhicules par analyse de photos. L'outil permet à la fois d'évaluer les dégâts et de détecter des fraudes éventuelles et de faire un premier devis estimatif des réparations. A la clé pour les assureurs, une réduction des coûts de l'expertise.

Enfin, les processus internes aux assurances peuvent être automatisés avec des solutions et méthode de « Robotics Process Automation » déjà évoquée au niveau des services financiers. Elles peuvent par exemple faire appel à **Captricity** (2011, USA, $52M) et à sa solution de gestion documentaire.

C'est aussi le cas de **Luko** (2016, France, 74M€), une assurance habitation qui avait 20 000 clients fin 2020 et qui automatise la gestion des sinistres avec des briques d'IA, en faisant notamment appel à l'agent conversationnel de la startup de paiement en ligne **Lydia** (2013, France, $159M). Elle propose par ailleurs des outils connectés (sécurité d'accès, détection de fuites d'eau, ...) de prévention des sinistres.

[1396] La source du schéma est Shaping the future of insurance with IBM Watson, novembre 2017 (43 slides).

Juridique

Le lancement de la startup **Ross Intelligence** (2014, Canada/USA, $13M) qui s'appuie sur IBM Watson il y a quelques années a créé un signal fort sur le marché : les métiers intellectuels comme celui d'avocat allaient être transformés radicalement par l'IA[1397].

Vus de près, une bonne partie des outils de l'IA dans les métiers juridiques sont des moteurs de recherche améliorés qui permettent de consulter la jurisprudence et les lois. Des applications plus élaborées ont vu le jour pour produire plus rapidement des contrats, les optimiser et pour générer des prévisions à bases probabilistes sur l'issue de procès[1398]. Les techniques d'IA juridique tournent principalement autour du traitement du langage et de la modélisation des connaissances. Comme pour les chatbots, elles sont encore imparfaites car le langage et le raisonnement sont très difficiles à modéliser et manipuler par des machines.

Ces solutions sont donc loin de remplacer les fonctions juridiques actuelles même si elles peuvent certainement améliorer la productivité des métiers du paralégal[1399] dans les entreprises et cabinets d'avocats, en particulier aux USA qui sont les plus gros consommateurs, de loin, de cette profession.

L'IA juridique comprend des outils qui améliorent la productivité de nombreux professionnels, comme les tableurs l'ont fait pour tous les métiers exploitant des données chiffrées depuis plus de 35 ans. Comme toute technologie qui se déploie largement, l'IA pourrait aussi permettre un élargissement du marché juridique tout en ayant un effet déflationniste sur certains tarifs. C'est le phénomène bien connu de la commoditisation.

[1397] Voir Legal Aspects of AI de Richard Kemp, novembre 2016, qui évoque à la fois les usages de l'IA dans les métiers juridique et les impacts juridiques de l'IA.

[1398] Voir ce bon panorama dans The Robot Lawyer? de Saskia Mehlhorn, 2017 (41 slides).

[1399] Voir l'excellent, long et très documenté Les robots, avocats et juges de demain ? Pas vraiment ... Intelligence artificielle en droit : derrière la "hype", la réalité par Emmanuel Barthe, mars 2018. D'une manière générale, les limitations des IA évoquées dans cet article sont liées au fait que la grande majorité des systèmes de recherche de texte juridiques exploitent des représentations mathématiques des textes dans des algorithmes de machine learning qui ne gèrent pas le sens des textes et n'ont pas de véritable système de représentation structuré des connaissances. Sauf si sont adjoints des moteurs de règles « à l'ancienne ».

Comme pour tous les marchés traditionnels, celui des services juridiques est affecté par les innovations proposées par des startups. Elles sont relativement nombreuses, aussi bien aux USA qu'en France[1400]. La cartographie *ci-dessus* en présente quelques-unes dont une partie seront détaillées dans la suite.

Dans **Legal Aspects of Artificial Intelligence**[1401] de Richard Kemp (novembre 2016), on trouve cette petite liste intéressante d'étude de cas d'usage de l'IA dans des cabinets d'avocats américains. Il s'agissait à l'époque d'effets d'annonce, sans véritables retours d'expérience. Depuis, les choses ont avancé et l'adoption de ces outils a fait son chemin, surtout aux USA.

Table - Examples of recently announced B2B AI use cases in the legal services market

Date	Law firm	AI provider	Use case
2015			
Aug	Dentons	IBM/ ROSS Intelligence	Dentons partners with IBM on IBM Cloud. Dentons' NextLaw Labs partners with Ross Intelligence to develop a legal app powered by IBM Watson[23]
Sept	Berwin Leighton Paisner (BLP)	RAVN Systems	RAVN Systems announces that BLP is using its AI platform to manage property Light Obstruction Notices[24]
Oct		Thomson Reuters/ IBM Watson	Thomson Reuters partners with IBM to deliver Watson cognitive computing solutions[25], with Legal as the first use case
Dec	Riverview Law	CIXILEX	Riverview launches the Kim Virtual Assistant built on the CIXILEX platform acquired by Riverview in August.
2016			
May	Baker Hostetler	ROSS Intelligence	Baker Hostetler becomes the first US law firm to license ROSS

Date	Law firm	AI provider	Use case
May	BLP	not stated	BLP wins the first contested High Court application to use Predictive Coding in litigation document disclosure[26]
May	Linklaters	RAVN	Linklaters confirms it has signed an MSA with RAVN[27]
June	Allen & Overy	Deloitte	Allen & Overy launch digital derivatives compliance system MarginMatrix with Deloitte[28]
June	DLA Piper	Kira Systems	DLA Piper announces agreement to use Kira in M&A due diligence[29]
July	Clifford Chance	Kira Systems	Clifford Chance announces AI agreement with Kira Systems[30]
Sept	Freshfields	Kira Systems	Freshfields announces agreement to use Kira in its Legal Services Centre[31]
Sept	Slaughter and May	Luminance	Slaughter and May announces collaboration with Luminance on legal due diligence AI[32]

Une étude de 2016 indique que la moitié des cabinets d'avocats US de plus de 1000 salariés utilisaient déjà des outils d'IA[1402]. Ce taux était inférieur à 10% pour les autres tailles de cabinets.

Cela rappelle la situation de nombreuses professions libérales (notaires, experts comptables, médecins) qui sont fragilisées par leur fragmentation face aux ruptures technologiques qu'elles sont lentes à adopter vis-à-vis de grandes organisations plus structurées. Reste à savoir à quelles applications les personnes sondées faisaient allusion, surtout dans la mesure où la définition de l'IA est à forte géométrie variable pour la plupart des professionnels de tous métiers.

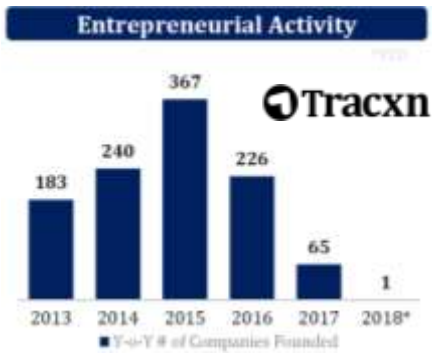
Le Rapport 2018 de **Tracxn** fait (*ci-dessous*) un point intéressant des domaines où les startups des legaltechs ont le plus levé de fonds, sachant qu'une bonne part utilisent de plus en plus des briques d'IA ans leurs solutions.

La gestion de contrats arrive en premier suivi des outils de eDiscovery qui permettent de préparer un dossier juridique de contentieux et comprennent les moteurs de recherche juridiques. La propriété intellectuelle arrive en quatrième.

[1400] Voir Legal tech report 2018 de la société d'analystes pour investisseurs en capital risque Tracxn, mai 2018 (214 pages) qui inventorie 1874 legaltechs dans le monde et cette Liste des legal techs françaises ou implantées en France de Benoît Charpentier datant de 2017 (6 pages) qui recense 94 startups legaltech en France, une partie seulement faisant appel à de l'IA. L'étude de Tracxn montre que le pic de création et de financement de legaltechs remonte à 2015, comme une indication que « les jeux sont faits » dans ce secteur.

[1401] Voir Legal Aspects of Artificial Intelligence, novembre 2016 (34 pages). Voir aussi L'intelligence artificielle au service de l'avocat : l'avocat-robot est-il à nos portes? d'Andrée-Anne Perras-Fortin et Eric Lavallée, septembre 2018, qui nous donne une perspective canadienne.

[1402] Voir Altman Weil Law Firms in Transition, 2017.



Key Sub Sectors in Legal Tech Sector

Sub Sectors	# Cos Tracked	# Funded	Total Funding
Legal Contract Management Companies offering contract lifecycle management platform which spans across contract...	240	70	$951M
eDiscovery Companies offering a platform to manage emails, documents, and media files for litigation...	130	52	$658M
Legal Documents/Forms Companies offering online DIY legal form services for individuals, SMBs, and Enterprises...	179	43	$377M
Intellectual Property (IP) Management Companies offering solutions for IP management including filing, tracking, analytics,...	174	59	$349M
Lawyer's Marketplace Companies offering legal services marketplace for connecting consumers, businesses, attorneys...	357	58	$203M
Law Enforcement Companies offering law enforcement software for law enforcement agencies	23	6	$130M
Legal Practice Management Companies offering legal practice management platform for end to end client intake to...	431	52	$101M

Moteurs de recherche juridiques

Une grande majorité des startups juridiques de l'IA proposent des moteurs de recherche dans les textes de lois et de jurisprudence. Elles fleurissent particulièrement bien dans les pays anglo-saxons dont le droit est influencé par la jurisprudence (dit de « case law »), tandis que le droit européen et surtout français, est plus fortement influencé par les lois et règlements (« civil law » ou droit romain).

Ross Intelligence (2014, USA/Canada, $13,1M), fondée par le canadien Andrew Arruda qui a fait une partie de ses études à la Sorbonne, s'appuie sur IBM Watson et est utilisé aussi bien par de grands cabinets d'avocats aux USA comme **Baker Hostetler** en mai 2016 ou petits comme **Salazar Jackson**[1403]. C'est essentiellement un moteur de recherche que l'on interroge avec des questions posées en langage naturel. Il est censé devenir plus intelligent au gré de son usage.

Ce n'est pas évident à comprendre car la réponse à des questions ne constitue pas une forme d'apprentissage supervisé ou par renforcement. Ross est en fait spécialisé dans les affaires de propriété intellectuelle et dans la gestion des faillites d'entreprises.

Il propose aussi son outil EVA qui analyse des documents juridiques comme des briefs d'avocats de parties adverses pour les annoter automatiquement avec des éléments de droit et de jurisprudence, permettant d'évaluer rapidement la situation (vidéo).

[1403] Voir Law Firm BakerHostetler Hires A 'Digital Attorney' Named ROSS, mai 2016 et Legal AI: It's not just for Big Law – Salazar Jackson and ROSS Intelligence, janvier 2017.

Nous avons divers autres systèmes d'interrogation de bases de connaissances de jurisprudence comme **Ravel Law** (2012, USA, $15,2M, acquis par LexisNexis en 2017), issu des écoles de droit et d'informatique de Stanford (vidéo).

Casetext (2013, USA, $20,8M) qui est focalisé sur l'analyse de jurisprudences avec sa solution CARA (Care Analysis Research Assistant) (vidéo).

Elle indique notamment les lois qui ont été le plus souvent cassées en jurisprudence, dans les pays de case law (vidéo).

Judicata (2012, USA, $7,8M) s'intéresse lui aussi à l'analyse de la jurisprudence. Ils détectent les propos déformés de manière délibérée par la partie adverse dans des procès.

Fastcase (1999, USA) propose un autre outil de recherche juridique (vidéo).

BloomBergLaw Points of Law est un moteur de recherche qui permet de préparer une plaidoirie en recherchant dans l'historique des opinions des juges. C'est encore une fois une solution adaptée uniquement au marché US.

Comme le droit est différent d'un pays à l'autre, les startups du secteur sont souvent cloisonnées au départ par pays. Très peu de startups couvrent à la fois les marchés juridiques des deux côtés de l'Atlantique.

Il y a aussi **Doctrine.fr** (2016, France, $12M)[1404] qui permet de faire des recherches en langage naturel dans une grande masse de bases de jurisprudence pour 129€ HT par mois (vidéo). Sachant qu'il n'y a pas beaucoup d'IA dans la solution. Elle sembler reposer en grande partie sur l'interrogation classique de bases de données.

Simulations et prévisions

Les outils de simulation et de prévision, souvent assimilés à la notion de justice prédictive, analysent des situations juridiques pour évaluer les risques et leur issue. Ils ne remplacent ni les avocats ni les juges, même si parfois, ils apportent un peu plus de rationnalité que ces derniers. Toutefois, cette rationnalité est à prendre avec des pincettes puisqu'elle s'appuie sur la jurisprudence qui est elle-même d'origine humaine ! Mais l'IA peu aussi servir à identifier des biais dans l'historique des décisions de justice[1405].

[1404] Voir Entre levée de fonds record et levée de bouclier de la profession, quel avenir pour Doctrine ? d'Anaïs Richardin, juin 2018, qui décrit les méthodes de growth hacking contestées de Doctrine.fr. La startup a été attaquée en justice par l'ordre des avocats du barreau de Paris en septembre 2018 pour usurpation d'identité, la méthode qu'aurait employé la startup pour récupérer des informations sur Infogreffe. Voir La start-up Doctrine attaquée par l'Ordre des avocats de Paris de Guillaume Bregeras, paru dans Les Echos le 27 septembre 2018.

[1405] Voir Setting a precedent in the use of artificial intelligence par Swinburne University of Technology, mai 2019. 200 paramètres entrent en ligne de compte dans les décisions de justice, dont des biais sur la couleur de la peau des prévenus, leur domiciliation et leur origine sociale.

Ceci étant, la perspective d'une justice prédictive exploitée par les juges eux-mêmes commence à effrayer[1406]. Avant même que cela fonctionne, elle inquiète déjà la profession et les colloques de juristes sur le sujet sont nombreux depuis deux ans.

Premonition (2014, USA, $5M) permet de sélectionner son avocat (aux USA) en fonction de son track record de gains de procès en fonction des juges qu'ils ont en face d'eux. La solution s'adresse aussi bien aux clients qu'aux avocats.

Case Law Analytics (2017, France) est une startup proposant une solution d'analyse des aléas juridiques d'une affaire. Là encore, il doit y avoir beaucoup de traitement du langage dans la solution. L'un des cofondateurs est un ancien directeur de recherche d'Inria spécialisé en modélisation mathématique en économie, finance et droit. La solution analyse la jurisprudence, établit et visualise des modèles probabilistes permettant d'évaluer non seulement l'issue d'une affaire mais également ses éléments quantitatifs comme les dommages et intérêts. Comme partout dans l'IA, la qualité d'une évaluation dépend de la quantité d'études de cas qui entraînent les modèles d'IA. Si votre cas est inédit, l'étude de la jurisprudence n'apportera pas grand-chose. Et elle est pertinente si la jurisprudence est abondante et bien commentée. Depuis fin 2017, le cabinet Flichy Granger spécialisé en droit social est équipé de Case Law Analytics[1407].

Predictice (2016, France) est aussi positionné dans la justice prédictive. C'est cependant une solution généraliste avec un moteur de recherche de documents juridiques. La startup exploite les données ouvertes de Légifrance (textes de droit) et Jurica (jurisprudence).

En matière de prévisions, le système fournit surtout une vue statistique et probabiliste des cas passés et les moyennes de dommages et intérêts. Les outils exploités intègrent l'analyse de langage et un moteur de règles.

[1406] Voir par exemple Un robot pour juge? Non merci par Miriam Mazou, août 2019.

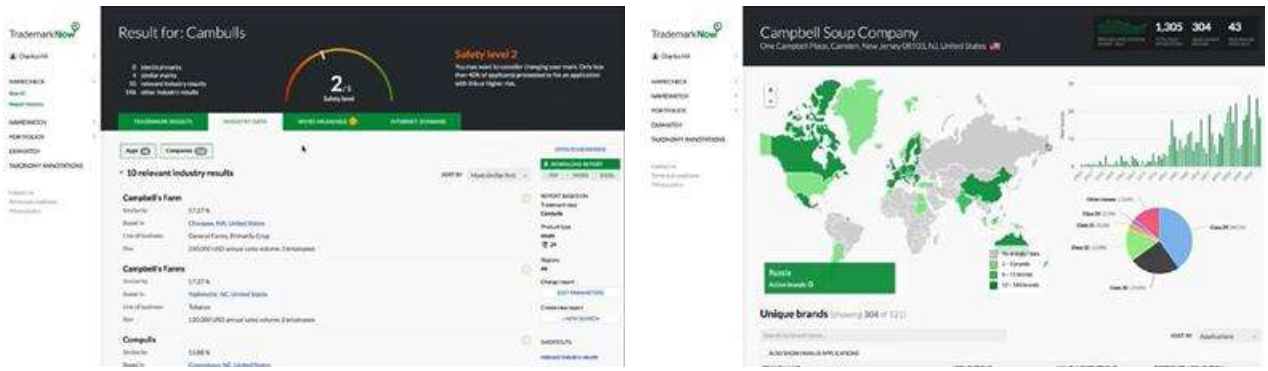
[1407] Voir Flichy Grangé Avocats utilise Case Law Analytics, un outil de justice prédictive, novembre 2017.

LexMachina (2009, USA, $10M, acquis par LexisNexis en 2015) fournit des outils d'analyse statistiques pour les avocats. Ces analyses sont plutôt quantitatives et temporelles. Elles permettent d'anticiper la longueur des procédures, la réaction des avocats de parties adverses et des juges et se s'organiser en conséquence. C'est une sorte de « Business Objects » pour avocats.

Propriété intellectuelle

La propriété intellectuelle est un champ à part entière du droit avec ses spécialistes, les conseils en propriété intellectuelle. Les bases de données associées sont spécifiques, à savoir celles des marques et des brevets.

- **Juristat** (2012, USA, $1,8M) réalise des analytics sur les données publiques sur les brevets et les avis des agents de l'USPTO pour optimiser les stratégies de protection de brevets. C'est une solution dédiée au marché US. Le financement de la startup l'explique en partie. Il est encore limité pour permettre un développement international rapide.
- **Turbo Patent** (2008, USA, $6,7M) produit des reportings sur la qualité et la valeur d'un portefeuille de brevets.
- **Lex Machina** (2009, USA, $10M, acquis par LexisNexis) fournit une solution de prévision sur les litiges de propriété intellectuelle. Elle exploite des solutions d'analyse du langage développées à l'Université de Stanford.
- **Data&Data** (2012, France) détecte en temps réel les ventes de contrefaçons et de marché gris sur Internet. L'outil à base d'IA s'appuie sur la détection d'anomalies dans les prix et les images des articles en vente.
- **Clarivate Analytics** (2016, USA) propose une large gamme d'outils de recherche dans la propriété intellectuelle couvrant brevets et marques, dans ce dernier cas, via l'acquisition de TrademarkVision qui utilise le deep learning pour reconnaître les logos de marques et faire des recherches d'antériorité, ainsi qu'une solution de lutte contre la contrefaçon issue de l'acquisition de MarkMonitor.
- **TrademarkNow** (2012, USA, $9,4M) propose une solution de recherche portant sur la protection des marques (*ci-dessous*). Il est cependant difficile d'y identifier des morceaux d'IA.



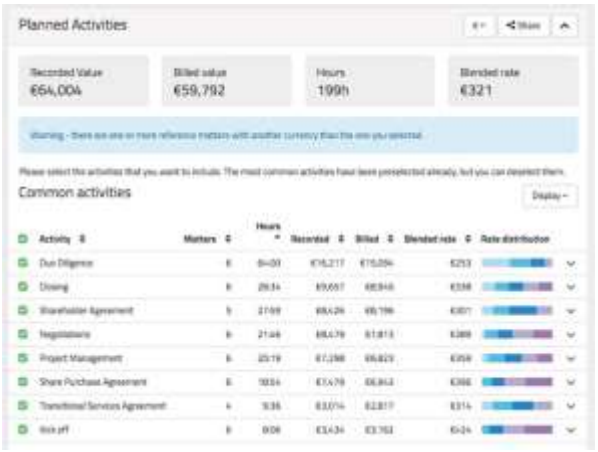
- L'office mondial des brevets, le **WIPO**, a mis en place en 2019 un moteur de recherche de marques et logos à base d'IA[1408]. C'est une application classique et somme toute évidente de la reconnaissance d'image à base de deep learning.

[1408] Voir WIPO launches AI-based trademark search, avril 2019.

Gestion de contrats

Nombre d'applications juridiques d'entreprises sont destinées à faciliter la gestion de contrats, pour identifier les clauses clés ou anormales, gérer les versions et faciliter le circuit de préparation et de signature. Les contrats commerciaux ou d'associations entre entreprises constituent en effet une grosse part du volume de travail des services juridiques des entreprises.

ClockTimizer (2014, Pays-Bas) propose un outil de business analytics pour cabinets d'avocats qui permet d'évaluer le temps passé sur des contrats clients et d'affiner ensuite les devis pour d'autres clients basés sur l'expérience. Les outils exploitent à la fois des données textuelles (mots clés des contrats, etc) et quantitatives (temps passé, etc).



Kira Systems (2015, Canada, $50M) propose une solution de « due diligence », de recherche, d'analyse et d'édition de contrats. Elle gère notamment les contrats de fusions-acquisitions ou de restructuration de capital d'entreprises.

C'est aussi l'activité de **eBrevia** (2012, Canada, $4,3M) qui a codéveloppé sa solution avec l'Université de Columbia qui couvre notamment le droit de l'immobilier et fournit des solutions d'analytics de contrats.

Hyperlex (2017, France, 1M€) est une plateforme en ligne de gestion de contrats. Elle analyse les contrats d'entreprise pour identifier les clauses clés et permet leur revue collaborative dans l'entreprise. L'outil comprend surtout un dashboard de suivi de contrats et de leur circuit de signature, une fonction de workflow assez traditionnelle.

LawGeex (2014, Israël, $21,5M) fait de la revue de contrats, notamment de confidentialité (NDA). Une étude montre que LawGeex est plus efficace qu'un avocat moyen dans l'évaluation de contrats de NDAs. En février 2018, leur solution obtenait un score de 94% de détection d'anomalies dans des NDAs contre 85% pour une vingtaine d'avocats humains[1409].

Icertis (2009, USA, $211M) propose une plateforme de gestion de contrats dans le cloud aux entreprises. Elle comporte des modules d'IA qui permettent d'ajouter une couche de traitement du langage et de recherche à leur solution. Elle sert notamment à valider les contrats tiers. Elle s'appuie sur une base de 5,7 millions de contrats dans 40 langues issus de 90 pays et 25 marchés. Ce qui pose la question de leur confidentialité au passage !

[1409] Voir Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts (37 pages) et LawGeex Hits 94% Accuracy in NDA Review vs 85% for Human Lawyers, février 2018.

Neota Logic (2010, USA) est spécialisé dans la préparation de contrats de confidentialité (NDAs). L'IA qu'elle contient est censée permettre la sélection des bons modèles en fonction des besoins. Ca peut être aussi bien un moteur de règles pas trop complexe ou un outil de machine learning exploitant quelques dizaines de variables et de la PCA (Principal Components Analysis) pour identifier les paramètres clés de choix des templates.

Klarity (2017, USA, $120K) gère la revue de contrats en cloud dans des fichiers Word qui sont automatiquement annotés. **Lexion** (USA, $4,2M) est un outil du même genre. **Legal Sifter** (2013, USA, $6,3M) est une autre solution d'analyse rapide de contrats qui détecte les points clés ou manquants.

Seal Software (2010, USA, $58M) associe un moteur de recherche juridique, Intelligent Content Analytics, exploitable lors de la gestion de contrats. La startup a fait l'acquisition d'Apogee Legal qui leur a permis de compléter cela avec des outils d'analytics de contrats pour la gestion des achats et la gestion de la vie privée ainsi que pour préparer les entreprises au Brexit. Elle était elle-même acquise par **DocuSign** en février 2020 pour $188M.

LinkSquares (2015, USA, $7M) est un autre moteur de recherche exploitable lors de l'édition de contrats. Le tout est géré dans le cloud.

Luminance (2016, UK, $23M) propose un outil d'édition de contrats issu de travaux de l'Université de Cambridge, qui lit et comprend les documents juridiques et exploite les interactions entre les avocats et les servcies paralégals et ces documents.

Autres usages

Au programme, nous avons notamment des outils de business analytics et des chatbots grand public.

On compte notamment **LegalZoom** (1999, USA), un service d'avocat en ligne couvrant à la fois le droit des affaires et le droit civil[1410] qui s'appuie sur IBM Watson.

Quid-IA (2018, France) propose des chatbots juridiques pour les entreprises et les particuliers. Ce serait le premier chatbot accompagnant les entreprises dans leur mise en conformité au RGPD. C'est en fait une filiale du cabinet juridique ALTIJ de Toulouse.

Cognigo (2016, Israël, $10,9M) aide également à gérer sa conformité au RGPD avec sa solution DataSense qui analyse de nombreux flux de données de l'entreprise. La cuisine interne de l'ensemble n'est cependant pas bien claire.

Ils indiquent surtout exploiter le traitement du langage pour identifier le type d'information qui circule, issue de données structurées et non structurées. Ils ont aussi l'air d'analyser la cybersécurité des données personnelles, notamment celles qui sont gérées dans le cloud. La startup a été acquise par NetApp en mai 2019.

DoNotPay (2016, UK) est un chatbot anglais créé par un jeune qui avait à l'époque 19 ans et qui permet de faire sauter les contraventions aux USA et au Royaume Uni et pour négocier la baisse de vos billets d'avions achetés, lorsque les prix baissent (on demande à voir…). Il a ensuite été étendu à la gestion de nombreux cas de droit civique. L'ensemble exploite IBM Watson, ce qui montre qu'avec un peu de détemination une personne isolée peut le paramétrer efficacement.

[1410] Voir le numéro d'août 2014 de The American Lawyer qui en parle bien.

Au passage, signalons que nombre de logiciels, les inventions à base d'IA sont brevetables si elles comportent des spécificités techniques nouvelles. Se pose la question de la brevetabilité d'une invention qui serait créée par une IA. On n'y est pas encore car la plupart des IA impliquées dans des inventions sont des outils exploités par des inventeurs homo-sapiens.

Aucune IA ne gère une invention de la tête aux pieds, de l'idée ou l'intention jusqu'à la formalisation et la mise en œuvre. C'est toujours un outil[1411].

AirHelp (2013, USA, $12,3M) a développé un « robot-avocat » dénommé Laraqui qui détermine les chances d'obtenir gain de cause en justice en fonction des éléments clés du dossier. Cela concerne en fait les réclamations dans le transport aérien.

Une autre startup, **RefundMyTicket** (2014, France) est aussi positionnée sur le créneau des remboursements de vols annulés ou retardés dans le transport aérien. Le site **Justice.cool** au sein du groupe **Claim Assistance** qui comprend Refundmyticket, évalue la légitimité des demandes par similarités avec un mix de machine learning et de moteur de règles.

Ravn Systems (2010, UK, acquise par iManage en 2017) a développé l'application LPP (Legal Professional Privilege) pour le Serious Fraud Office, l'organisme de l'état Britannique qui gère les grandes affaires de criminalité financière, fraudes et corruption, une sorte de Tracfin étendu. LPP leur sert à passer au peigne fin les documents juridiques des affaires en cours.

Atrium LTS (2017, USA, $75M) est un cabinet d'avocats numérisé s'appuyant sur diverses formes d'IA non précisées pour automatiser son activité, mais pas encore à 100%. Ils visent notamment les startups (vidéo).

Justice.cool (2020, France) est une plateforme de médiation juridique assistée par une IA qui permet de régler rapidement les petits litiges en ligne dans le civil aussi bien que dans les relations employés/employeurs. La solution analyse le contexte des affaires et les dispositions prévues par la loi pour fournir aux parties prenantes des indicateurs statistiques. Cela permet de prendre des décisions en connaissance de cause. Le tout exploite surtout du traitement du langage qui analyse le litige et identifie les demandes prévues par la loi. L'IA gère ensuite le processus du règlement probablement en mélangeant des techniques classiques et d'autres qui relèvent véritablement de l'IA.

Distribution

La distribution est un terrain propice au déploiement d'applications exploitant diverses briques d'IA[1412]. Que ce soit en ligne ou dans la distribution physique, les acteurs de ce marché disposent de très gros volumes de données : les catalogues produits, les inventaires, les flux de vente ainsi que les achats et habitudes des clients.

Des contrastes demeurent entre le commerce en ligne et le commerce physique traditionnel. Vu de loin, les premiers ont plus numérisé leurs processus que les seconds. Mais dans les deux cas, les acteurs de ce marché peuvent accumuler des tombereaux de données, surtout d'ailleurs pour les enseignes qui sont à la fois en ligne et dans la distribution physique. Dans les deux cas, les capteurs et l'IA permettent de pister les clients, de la navigation dans un site web à la déambulation dans un magasin et dans ses rayons.

[1411] Voir Les robots, des inventeurs comme les autres ? de Magali Touroude, avril 2018.

[1412] Voici quelques sources d'informations où j'ai pu glaner quelques éléments sur l'IA dans le retail : Artificial intelligence for High Frequency Retail – Pricing, Inventory and Margins Optimization, septembre 2018, The retail renaissance: Leading brands use data and AI to win, avril 2018, Artificial Intelligence in Retail, Part 1: Applications Across Customer-Facing Functions de Coresight Research, 2018 (12 pages), Artificial Intelligence In The Retail Industry de Shaily Kumar, novembre 2017 et AI retail playbook de Microsoft (29 slides). Voir enfin The AI Industry Series: Top Retail AI Trends To Watch - CB Insights Research, 2019, qui fait un bon panorama de solutions pour le retail à base d'IA. La page permet de récupérer un bon rapport de 40 pages en échange d'un mail professionnel.

Les données permettent aussi de mettre en place des techniques d'optimisation du pricing à base de machine learning, un peu comme le yield management des les compagnies aériennes[1413].

Les innovateurs du secteur passent leur temps à inventer des solutions technologiques pour prédire les besoins des clients et améliorer les manières de leur mettre sous le nez ce qu'ils pourraient acheter. Si possible, avec des achats impulsifs. Pourtant, les clients ont des besoins assez basiques dans nombre de catégories produits : trouver rapidement ce qu'ils cherchent, pouvoir comparer les offres et trouver les meilleures au niveau fonctionnel et tarifaire, pouvoir obtenir le produit rapidement, et pouvoir le dépanner ou retourner facilement en cas de problème pour les produits non consommables.

L'IA ne peut pas s'intégrer dans une méthode Coué justifiant d'éviter la perspective désagréable des crises à venir pour le retail physique.

Celle-ci a surtout besoin de rendre ses points de vente attirant visuellement, émotionnellement et dans la qualité de service, pour éviter l'inexorabilité de la fuite vers le commerce en ligne. L'IA peut y aider, mais elle n'est pas la seule en jeux. Il ne faut jamais perdre la dimension humaine et bien interpréter les bouleversements sociétaux des modes de consommation, de transports, de relation au temps et de relations interpersonnelles.

Les besoins des commercants ? Ce sont des intermédiaires entre les marques et les consommateurs. Ils doivent analyser les tendances, comprendre les sentiments autour des marques, mettre les produits bien marketés dans les maisons des consommateurs, faire de l'upselling et du cross-selling, réduire leurs frais de gestion, optimiser les stocks et leur rotation, limiter la fraude (en ligne) et la démarque inconnue (en magasin). Les retailers veulent soit prédire le futur, soit l'influencer à leur avantage.

Nous allons ici faire le tour de quelques cas d'usages et de startups de l'IA dans le retail en séparant le retail physique du commerce en ligne[1414], sachant que cette frontière est de plus en plus floue entre les deux, avec le *phygital* et le web-to-store et autres variantes du genre.

Distribution physique

L'IA permet de répondre aux besoins de retailers traditionnels côté assortiment produit, optimisation des linéaires, recommandation, dans le web-to-store, le paiement et la lutte contre la démarque inconnue.

Recommandation

La recommandation de produits, qu'elle qu'en soit la forme, est le Graal du retail et de l'IA. Il n'est donc pas étonnant que ce soit dans cette catégorie qu'agissent le plus grande nombre de startups du secteur. Cette recommandation intervient généralement sur applications mobiles, soit en amont des achats soit en magasin. Elle exploite des outils de machine learning qui s'appuient sur des historiques de vente, des données géolocalisées et autres moyens de profiler les clients et prospects. Dans la mobilité, la publicité et la recommandation peuvent apparaître dans l'application du retailer, dans des publicités distribuées par des régies traditionnelles ou être distribuées tout simplement par SMS ou via les réseaux sociaux.

[1413] Voir 10 Ways AI Improves Pricing And Revenue Management par Louis Columbus, septembre 2020 qui fait un très bel inventaire des techniques d'optimisation des prix et des revenus.

[1414] Dans Why retail's artificial intelligence bet is all wrong, mars 2018, Luis Perez-Breva calme les ardeurs du "tout IA" dans le retail en mettant en avant le fait que l'IA doit aider les humains, pas les remplacer, notamment dans les points de vente. Il rappelle aussi que les retailers disposent de moins de données que les e-commerçants sur le profil des acheteurs, tout du moins dans la distribution spécialisée. En France, un hypermarché sait beaucoup de choses sur vous, même s'il n'exploite pas bien ces données. Il avance aussi que la recommandation est survendue, y compris dans le cas d'Amazon. Il la juge faiblement pertinente mais attractive pour les clients car cela les occupe. Il relève aussi un point clé que je met en avant depuis quelques années : le fait que le logiciel et l'IA transfère du travail rémunéré des fournisseurs vers du travail non rémunéré chez les clients.

Occi (2015, France) permet aux commercants d'envoyer des recommandations personnalisées à leurs clients sur leur smartphone, pour peu qu'ils les aient convaincus d'installer leur application ou qu'ils disposent de leurs numéros de téléphone ou de leurs identifiants de réseaux sociaux.

Reflektion (2012, USA, $42,8M) aide les commerçant à convertir les prospects en clients à partir de son moteur de recherche de point de vente pour les grandes enseignes comme Disney.

Armis (2015, France, 7M€) s'appuie sur le FLAI (Fast Learning Artificial Intelligence), une expression de leur cru, pour générer de la publicité promotionnelle en ligne optimisée pour chaque magasin visité par les clients. C'est testé ou déployé chez Intermarché et Bricorama.

S4M (2011, France, $20M) propose Fusio, une plateforme de publicité mobile dite « de drive-to-store ». Les publicités sont personnalisées. Ils commercialisent leur offre sur la base du Cost Per Incremental Visits (CPIV) ou du Cost Per Landing Page (CPLP) qui permet d'éviter les effets de la fraude à la publicité en ligne.

Teemo (2014, France, $17,9M), anciennement Databerries, est un spécialiste du ciblage de clients mobiles. Il les cible en fonction des lieux qu'ils visitent, par triangulation des signaux Wi-Fi, si celui-ci est activé dans les smartphones. Ils ont déjà plus d'une centaine de clients dont Leroy Merlin, Jardiland, Carrefour et Casino. Le tout pour pousser promotions et recommandations.

"Real Life Targeting"
cible les mobinautes en fonction des lieux qu'ils ont visités et mesure l'impact sur le trafic généré en magasin.
>100 clients dans l'automobile, l'alimentaire, le bricolage et l'ameublement (Volkswagen, Carrefour, Brico Dépôt, Gautier)
startup française, ex Databerries
levé $17,9M

Rubikloud (2013, Canada, $45,5M) promet monts et merveilles aux retailers. En gros, ils alimentent leur système avec toutes leurs données, et il va vous proposer les promotions à lancer pour optimiser les ventes, prévoir les ventes et les résultats de vos campagnes marketing et les augmenter de 10%. C'est mis en œuvre à Hong Kong, donc un peu loin pour vérifier ce qu'il en est !

Singlespot (2015, France, 6M€) développe une solution de publicité géolocalisée et ciblée sur mobiles permettant aux commerçants de communiquer de manière ciblée. Comment la collecte des informations se fait-elle ? Par des partenariats avec des éditeurs d'applications mobiles. Le tout en respectant le RGPD.

Untienots (2015, France, 1,7M€) personnalise les programmes de fidélité des distributeurs en exploitant les données des cartes de fidélité, la géolocalisation des clients et leur historique de navigation. Capté comment ? Via un logiciel mobile de gamification (*ci-contre*). Le tout, pour faire des recommandations ciblées (vidéo). La solution serait utilisée chez Auchan et Carrefour. Je suis client d'Auchan depuis des décennies et utilisateur de leur carte de fidélité mais je n'ai rien vu. Ce sont les affres du marketing !

TwentyBN (2015, Allemagne, $12,5M) a créé une solution d'affichage dynamique qui exploite un avatar affiché sur un grand écran. Celui-ci interagit avec les clients en discutant avec eux et en reconnaissant leur posture. Avec une déclinaison de recommandation dans le fashion (vidéo). Ils déclinent cela pour entraîner les clients dans les salles de sport sous l'appellation Millie Fit.

Target2Sell (2012, France, 5,5M€) utilise le machine learning pour faire de la recommandation personnalisée dans l'expérience web-to-store proposée aux clients des enseignes. La solution est déployée en France et à l'étranger chez des enseignes telles qu'Auchan, Sephora, Marionnaud, Yves Rocher, Cartier, But, Manutan et Lapeyre.

Advalo (2014, France, 5M€) propose une solution de marketing prédictif à base de machine learning ciblant la distribution spécialisée (équipement de la maison, habillement) mais aussi les concessionnaires automobiles qui aide à détecter en ligne les « prospects chauds » à la recherche de véhicule et les clients déjà connus, et les données associées (CRM, autres), qui seraient en phase de renouvellement. L'outil alimente en prospect les commerciaux sédentaires des concessionnaires.

Reveal Systems (2005, France) lançait fin 2019 AR Navigation, une application mobile de navigation dans les centres commerciaux associant une fonction de navigation précise et de la réalité augmentée n'utilisant ni GPS ni le WiFi. C'est la fonction photo du smartphone qui permet à l'application de reconnaître l'emplacement dans le centre commercial qui a été auparavant cartographié en 3D. On est ici dans le champ de la vision artificielle.

Efin, **Ysance** (2005, France, 5M€) collecte et analyse les données des consommateurs et prospects pour prédire leur parcours d'achat en mode phygital (retail+online). Ils sont présents dans 80 pays et ont notamment L'Occitane, Etam, la Société Générale et Fnac-Darty comme clients. Reste à savoir comment tout cela est mis en musique et avec quelles données exactement[1415]. La société a l'air de fournir du « service outillé » avec des briques en propre qu'elle adapte en fonction des besoins des clients et de leur système d'information.

Emplacement

Le machine learning peut être mis à contribution pour aider aux choix d'un emplacement d'un nouveau point de vente. De nombreuses sources de données plus ou moins ouvertes peuvent être exploitées : les données de cadastre, les données de circulation des mobiles provenant des opérateurs télécoms et autres données de trafic routier ou de transports en commun.

symaps.io (2016, France, $700K) aide ainsi à identifier le bon emplacement pour un point de vente avec un mélange de technologies de machine learning et de géomapping.

Optimisation des linéaires

L'usage de caméras vidéo est de plus en plus courant pour comprendre le comportement des clients dans les points de vente et optimiser l'organisation des linéaires ou celle des vendeurs, selon le type de produits distribués.

Percolata (2012, USA, $14,7M) exploite caméras de surveillance, captation audio, détection de smartphones et machine learning pour prédire le trafic dans les magasins. Il croise ces données avec l'historique de performance des vendeurs pour planifier les équipes de vente générant le plus haut niveau de chiffre d'affaire.

Angus.AI (2014, France, $500K) propose une solution de gestion des linéaires dans la distribution qui s'appuie sur des caméras de surveillance et celles des smartphones. Elle permet de gérer plus efficacement le réassortiment des rayons et le contrôle des prix. La solutions serait évaluée ou déployée chez Carrefour et Leroy Merlin[1416].

Predictix (2005, USA, $40M) est une startup créée par des Tunisiens qui fait partie du groupe américain Infor depuis 2016. C'est un spécialiste de l'optimisation de linéaires. Comme son nom l'indique, elle doit utiliser des techniques de machine learning pour faire du prédictif.

Celect (2013, USA, $15,2M, acquis par Nike en août 2019) permet aussi d'optimiser l'assortiment des rayons en fonction d'analyses prédictives comportementales des clients.

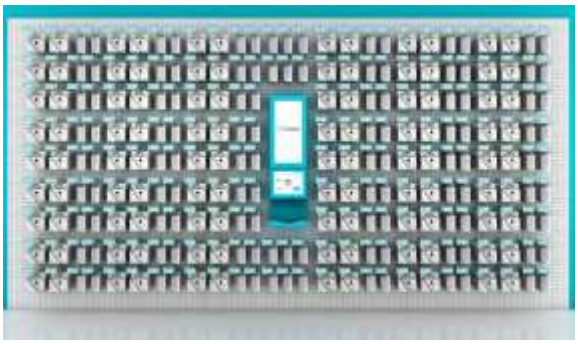
[1415] Voir Ysance, l'IA qui a séduit les plus grands retailers par Mathilde Degorce, 2018.

[1416] La startup avait été créée par des anciens ingénieurs d'Aldebaran qui ont développé la partie logicielle des robots Nao et Pepper. Leur première offre était une solution logicielle embarquée dans les robots leur apportant les fonctions de base de reconnaissance vocale et faciale et de détection d'obstacles. Le tout fourni sous la forme d'un SDK et d'APIs (interfaces de programmation). Ils s'appuient beaucoup sur des solutions open source du marché. Le marché du retail leur a ensuite semblé plus prometteur, et hop, un pivot !

Vekia (2008, France, 14,4M€) utilise le machine learning pour faciliter l'optimisation d'approvisionnement (vidéo). Sa solution est déployée chez Leroy Merlin, But, MrBricolage et Jacadi.

Yoobic (2014, UK, $30M) optimise la gestion des points de vente du retail. Il permet la communication entre les marques et leurs revendeurs ou entre franchises et franchisés. L'outil sert à la formation des équipes, à accéder à la documentation sur les produits, à gérer les stocks et à l'affichage dynamique dans les magasins. En gros, cela permet aux marques de vérifier que le marchandizing de leurs produits est bien réalisé dans les points de vente. C'est censé utiliser du machine learning, mais sans que cela soit bien précisé. Cette startup crée par des Français s'est installée à Londres. Ils comptent notamment The Kooples, Lacoste, Ladurée, Celio, Etam, Fnac-Darty et Casino comme clients.

SES Imotag (Allemagne) a fait l'acquisition de la startup **Findbox** (2012, Allemagne) qui a créé une application mobile permettant de retrouver un objet dans les rayons dont on dispose d'une photo. L'application est reliée à des dispositifs de fléchage du produit dans les rayons exploitant des LEDs. L'objet était à l'origine capté par un scanner 3D intégré dans un dispositif mural installé chez le retailer. Il a été depuis remplacé par une application mobile et de la reconnaissance d'image.



Enfin, **RoofStreet** (2015, France, 1M€) analyse et anticipe les déplacements des personnes pour leur faire ensuite des propositions adaptées d'implantations de points de vente et de l'analyse de concurrence, bref du géomarketing prédictif.

Prévisions

Lokad (2008, France) est un spécialiste de la *supply chain* pour le retail. Ils aident les commerçants à prévoir ce qui va leur arriver, notamment au niveau de la demande client, afin de mieux planifier les tâches de réassortiment. Ce sont de gros utilisateurs de machine learning. Ils ciblent aussi les entreprises des secteurs industriels pour leur permettre d'optimiser leur propre supply chain. Ils réussissaient en 2020 à être classé 6e dans un challenge international de prévision de ce genre[1417].

IBM propose une solution d'analyse des données clients et de sources diverses pour anticiper les besoins du marché et adapter les inventaires et les stratégies de tarification. IBM propose aussi un **Personal Shopper** été réalisé en partenariat avec **Fluid** (1999, USA, $24M, acquis par Astound Commerce en 2018). Le premier client est la chaîne de distribution de vêtements sportifs **North Face**. Il s'agit aussi d'un agent conversationnel utilisable via le service en ligne du site marchand.

Le corpus de données utilisé exploite tout le catalogue du site ainsi que les différents critères de choix des vêtements. Le dialogue proposé est très "scripté". Son arborescence semble limitée. Le système a été présenté au Big Show 2016 de la National Retail Foundation à New York[1418].

[1417] Voir Ranked 6th out of 909 teams in the M5 forecasting competition par Johannes Vermodel de Lokad, mai 2020.

[1418] Pour en savoir plus voir ce compte-rendu détaillé sur le JDN : Comment The North Face a appliqué Watson à l'expérience d'achat, de Flore Fauconnier, janvier 2016.

Analyse du visitorat

Trax (2010, Singapour, $386M) utilise la vision artificielle pour analyser les rayonnages et le positionnement des produits. L'outil est employé par un visiteur professionnel du point de vente.

Quidivi (2006, France, $1,4M) est une startup française qui analyse le visitorat en magasin par la vision artificielle. Ses outils mesurent non seulement le trafic mais aussi l'attention. Ils détectent l'âge et le sexe des visiteurs. Ancienne dans l'écosystème, la startup a fait évoluer ses techniques d'analyse d'image au gré des évolutions technologiques du deep learning. **Modcam** (2013, Suède, $7,6M) est un concurrent.

Paiement

Diverses solutions apparaîssent pour automatiser le passage en caisse et le paiement en boutique, comme avec **Aipoly** (2015, USA) qui utilise des caméras pour détecter les produits achetés (vidéo). Je les avais découverts au CES 2018.

Standard Cognition (2017, USA, $86M) (vidéo), **Zippin** (2018, USA) (vidéo) et **Inokyo** (2018, USA) (vidéo) proposent tous la même solution permettant aux clients d'un supermarché d'être automatiquement facturés via la reconnaissance des produits qu'ils prennent dans les rayons. Cela repose sur des caméras installées au-dessus des allées ou dans les rayons selon les cas et les configurations.

Wynd (2013, France, $123,5M) a développé une solution omnicanale de vente en ligne adaptée au retail classique pour unifier les canaux offline et online, qui intègre la logistique de livraison et l'encaissement. Elle permet de livrer les clients du retail classique un peu n'importe où. Ils sont notamment exploités par Carrefour pour son service de livraison express.

Cette tendance à créer des magasins avec checkout automatique a été lancée par **Amazon** avec une boutique Go, inaugurée à Seattle en décembre 2018, puis par **Wallmart** qui testait un concept voisin au printemps 2019[1419].

Caper (2017, USA, $13M) a créé un caddie qui scanne tout seul les produits que l'on place dedans grâce à plusieurs caméras. Il utilise dont au minimum du deep learning de reconnaissance d'images associé à un modèle bien entraîné avec tous les produits vendus dans le point de vente. Un écran affiche les caractéristiques et le prix du produit qui vient d'être scanné et peut faire de la recommandation d'upselling ou de cross-selling au consommateur. Un terminal de paiement est également intégré dans le caddie. Ca a l'air bien sur le papier à ceci près que le caddie coûte cher et est fragile. Pas sûr que cela soit facile à généraliser avec le vandalisme ambiant (vidéo)[1420].

Enfin, après le paiement, **Revers.io** (2009, France, $5,9M) optimise la gestion des retours dans le retail en gérant l'ensemble du processus et de la relation client associée, impression de bon d'envoi compris, le tout avec une solution logicielle en cloud (vidéo). C'est notamment utilisé à la Fnac.

[1419] Voir Walmart's AI-based store concept is open to the public par Jon Fingas, avril 2019.

[1420] Voir Meet Caper, the AI self-checkout shopping cart par Josh Constine, janvier 2019.

Fraude

StopLift (USA) est utilisé pour détecter la fraude mais pour les caisses automatiques self-service. Toujours avec l'emploi de caméras de surveillance couplées au système d'encaissement pour vérifier que les produits sont bien tous scannés et pesés. Le produit a été développé au sein du MIT.

Everseen (2007, Irlande, $13,5M) a créé timi.ai, qui détecte la fraude en sortie de caisse dans les points de vente via l'usage de caméras de surveillance.

Vaak (2017, Japon, $500K) utilise aussi la reconnaissance d'images mais pour détecter les vols à la tire dans les magasins en surveillant les clients dans les rayons[1421].

Commerce en ligne

Le commerce en ligne intègre évidemment presque tout ce que l'on trouve de nouveau dans le marketing, pour le ciblage publicitaire et commercial, sur les techniques d'upselling et cross-selling basées sur la recommandation qui s'appuient sur le machine learning ou dans la relation clients, chatbots compris. Le commerce en ligne peut exploiter quelques autres fonctionnalités à base d'IA que voici.

Optimisation du parcours client en ligne

Avec **Gainsight** (2011, USA, $156M), **Jetlore** (2011, $10,6M, acquis par Paypal en 2018) et **OnCorps** (2011, USA, $2,3M). **Granify** (2011, Canada, $13M) va jusqu'à suivre pas à pas le parcours en ligne des clients pour détecter ceux qui pourraient abandonner le panier en cours de constitution et leur proposer une action ou information permettant de l'éviter.

Influans (2016, France, 6M€) optimise aussi le parcours client pour leur proposer le bon produit avec les bonnes incitations et au bon moment.

TargetToSell (2012, France, 5M€) s'appuie sur un mix de machine learning classique et d'un réseau de neurones pour optimiser le parcours du parcours en fonction du profil du visiteur et de ses goûts de produits captés au niveau du site, cela va jusqu'à personnaliser le bandeau du site en fonction du profil de l'utilisateur.

Nextuser (2013, USA, $2,3M) se positionne comme un outil d'intégration des outils marketing. Enfin, **Perfect Path** (Canada) propose aussi un outil de modélisation du parcours client qui sert à définir les meilleurs scénarios de transformation en fonction des objectifs et segments visés.

Levia (2018, France) propose un agent conversationnel aux retailers suivant tout le parcours consommateur. Il transforme notamment des bases de connaissances en agent conversationel à même de répondre à des demandes complexes en langage naturel. La qualité doit bien entendu dépendre, entre autres choses, de la qualité des bases de connaissances. Le système exploite notamment les modèles Bert, CamemBert et GPT2 / GPT3. Ils sont notamment utilisés par Rueducommerce et Cdiscount. La société a été créée par les serial-entrepreneuses Lara Rouyres et Tatiana Jama.

Recommandation

On en trouve un peu partout comme chez **Turi** (2013, $23,5M, acquis par Apple en 2016), une startup montée par des anciens de Carnegie Mellon sous la forme initiale d'un projet open source. Même histoire avec **Reflektion** (2012, USA, $42,8M), adopté par Disney et Converse, qui propose du ciblage produit temps réel.

[1421] Voir Deep Science AI joins Defendry to automatically detect crimes on camera par Devin Coldewey, mars 2019.

Kameleoon (2008, France, 2,3M€) fait de la personnalisation et de la recommandation ciblée aux visiteurs d'un site web à base de machine learning. La solution est déployée chez Cdiscount et Allopneux.

Skapânê (2015, France, 500K€) propose diverses solutions exploitant le machine learning pour faire des prévisions en temps réel pour la recommandation de produits, la maintenance prédictive et la lutte contre la fraude. Basés à Lille, ils ont notamment pour client Auchan.

Paylead (2016, France, $671K) améliore la connaissance des clients en analysant l'historique de leurs transactions et comportements d'achat. Leur solution permet de fournir des recommandations pour des offres spéciales, achats et programmes de cashback.

Segmentify (2015, Turquie, 1M€) fait du marketing en ligne par email et application mobile pour harceler les clients et leur "*proposer les produits qu'ils aiment*". Evidemment, cela s'appuie sur du machine learning. Bref, c'est surtout un outil de recommandation.

Amazon StyleSnap permet à partir d'une photo de retrouver dans le catalogue d'Amazon les vêtements photographiés portés par n'importe qui, une fonctionnalité intégrée dans l'application mobile d'Amazon. Elle exploite notamment des réseaux convolutifs de type ResNet[1422].

Endor (Israël) fait de la « social physics », un autre nom pour la psychohistoire, cette pseudo-science inventée par Isaac Asimov, qui permet de prévoir à partir de signaux faibles ce que nous achèterons le mois prochain. Enfin, peut-être !

Enfin, **Nuukik** (2013, France) fait de la recommandation personnalisée aux consommateurs pour les sites de vente en ligne.

Commerciaux sédentaires

L'optimisation de leur activité et la prévision du comportement des clients avec des startups particulièrement bien financées est proposée par **InsideSales** (2004, USA, $251M), **Clari** (2012, USA, $61M), **Wise.io** (2012, USA, $3,6M, acquis par GE en 2016) et **Spiro** (2014, USA, $4,5M). Il est assez difficile de départager toutes ces belles startups !

Gestion d'offres produits

Akeneo (2013, France, $61,4M) propose une solution open source de PIM (Product Information Manager) pour les marketeurs qui permet de gérer ses catalogues produits en exploitant des données structurées variées (vidéo). La nouvelle V3 intègre un moteur d'IA dénommé Franklin qui permet d'accéder à une base de données de 50 millions de produits.

Futurescore.IO (2017, France) permet de réaliser des enquêtes clients sur une application mobile et d'exploiter une IA maison pour mesurer et localiser le potentiel commercial des produits, plutôt dans le fashion. Du machine learning est utilisé dans la solution, sans plus de précisions.

Le fashion peut faire appel à des solutions permettant de trouver la bonne taille et le bon style pour s'habiller avec **Thread** (2012, UK, $16,32M), le styliste en ligne **Stitch Fix** (2011, USA, $79,4M) qui exploite du dialogue en langage naturel, **Volumental** (2012, Suède, $4,5M) pour le choix de ses chaussures et **Thirdlove** (2013, USA, $13,6M) pour déterminer la taille de son soutien-gorge. Histoire de compenser le fait que les clients ne connaissent pas toujours leur taille.

[1422] Voir StyleSnap will change the way you shop, forever par Arun Krishnan, 2019.

Le benchmark des prix et des promotions des concurrents est réalisable avec **Daco** (2016, France) qui s'appuie sur du deep learning et de la reconnaissance d'images et fait du prédictif sur les actions des concurrents, le tout avec une couverture mondiale. Ils ont notamment comme clients L'Oréal, Vente-privée et les Galeries Fafayette.

Enfin, **Brennus Analytics** (2015, France) est focalisé sur les ventes b2b et peut ajuster les prix de manière dynamique avec des techniques de machine learning multicritères. Leur solution analyse, compare et prédit les comportements d'achat des clients pour adapter les prix à la stratégie marketing et commerciale comme les gains de parts de marché, de marge tout en prenant en compte les contraintes comme la charge des usines, des stocks.

Recherche visuelle

Il s'agit ici de permettre aux clients de retrouver un produit via sa photo ou des outils de recherche visuelle.

Ce domaine comprend les offres de **ViSenze** (2012, Singapour, $14M), **Cortexica** (2008, UK, $9,2M) et son logiciel findSimilar en cloud, **Grokstyle** (2016, USA, $2,5M), **Syte** (2015, Israël, $31,6M), qui est utilisé chez Conforama depuis 2019, et **Slyce** (2012, USA, $37M de levés, IPO en avril 2015) qui fait de la reconnaissance d'objets dans les applications mobiles, notamment pour Ikea.

Le moteur de recherche FashionBot de **GoFind** (2016, USA) permet de retrouver dans un site en ligne ce que l'on trouve dans un magasin[1423].

Twiggle (2014, Israël, $35M) offre un outil d'optimisation des recherches textuelles qui imite le comportement d'un commercial traditionnel. Enfin, ce genre de service est aussi proposé par **Pinterest** avec sa fonction Shop the Look (vidéo). La recherche visuelle est aussi l'une des principales fonctions de **Google** Lens.

Fraude et contrefaçon

Skymind (2014, USA, $6,3M), crée par des anciens de Vicarious, propose une solution open source en Java – Deeplearning4j.org – capable d'analyser des flux de données. Elle est notamment utilisée dans la détection de fraude, le commerce et le CRM.

[1423] Voir Visual Search - The Ultimate Guide: Statistics, News, Trends, and Tips par Clark Boyd, novembre 2019, qui fait un bon inventaire des cas d'usage de recherche visuelle qui ont plutôt court dans le retail physique. Notamment chez Walmart, Conforama, IKEA, Wayfair (électroménager),Marks & Spencer, Neiman Marcus (grands magasins aux USA). En plus des usages chez Pinterest et eBay. Voir aussi Shopping With Your Camera: Visual Image Search Meets E-Commerce at Alibaba, août 2018 qui fait référence à Visual Search at Alibaba, 2018 (9 pages) et vidéo, au sujet de l'application mobile Pailitao qui permet de retrouver les produits dans le service mobile Taobao en Chine, une filiale d'Alibaba. Son principe est de photographier un objet physique et de le retrouver ensuite dans l'offre en ligne.

Fraud.net (2015, USA) fait de la détection de fraude et s'appuie probablement sur du machine learning.

Data & Data (2012, France) analyse les ventes de produits contrefaits et marchés gris sur Internet (sites de e-commerce, places de marché, réseaux sociaux comme Facebook et Instagram). Cela permet notamment de vérifier la fidélité de son réseau de revendeurs autorisés.

Relation client

Elle fait appel à des chatbots en tout genre avec l'américain **Satisfi Labs** (2016, USA, $1M) et l'agence **TheChatbotFactory** (2012, France) qui créé des chatbots sur mesure. Elle a créé un chatbot sommelier pour Auchan sous Facebook Messenger.

La Redoute a déployé début 2018 un agent vocal dans son application mobile qui est associé à un outil de reconnaissance visuelle de produits photographiés avec son smartphone.

Eptica (2001, France, $13,4M, acquis par le Canadien **Enghouse Systems** en octobre 2019) propose une solution de création de chatbots pour la relation client.

Clustaar (2013, France, 1,7M€) permet de développer des chatbots de support client.

Ces chatbots sont complétés par des outils de gestion des flux de mails clients comme chez **Alcmeon** (2011, France) et son outil de workflow permettant de distribuer de manière semi-automatisée les sollicitations des clients aux équipes internes. Il est complété par un chatbot. Le tout est notamment utilisé chez Carrefour, Oui.sncf et Orange.

Hubware (2016, France, 1,4M€) gère les messages clients entrants (emails, réseaux sociaux…), les classifie, les complète d'informations de contexte et les enrichit de propositions de réponses qui sont alors exploitées par les agents du service client qui restent en frontal client. Ils fournissent notamment EDF.

Cognitive Matchbox (2016, France) améliore l'aiguillage des clients vers les bons agents dans les centres d'appels en fonction des émotions et des besoins. Cela rappelle ce que fait Batvoice. Ca tombe bien puisque les clients aiment bien tomber sur de vraies personnes lorsqu'ils appellent une société et pas sur un robot qui va leur répéter ce qu'ils ont déjà vu sur leur site web !

Logistique

La logistique du commerce et plus généralement la « supply chain[1424] » peuvent exploiter à de nombreux endroits des briques d'IA[1425] :

- La **détection des tendances** dans les réseaux sociaux qui permettent de prédire des ventes soudaines et l'engorgement de chaines logistiques comme les toupies à main en 2017.
- La **robotique de tri et de dispatch de colis** dans les entrepôts de logisticiens, comme avec **Exotec Solutions** (2015, France, $111M) est ses robots de transports de produits associés à des postes de préparations de commandes[1426]. Amazon est évidemment assez friand de ce genre de solution.

- L'**inspection visuelle d'entrepôts** comme avec **Qopius** (2015, France, 1,8M€) qui passe par des robots dotés de fonction de vision artificielle relativement classique.

  Cette inspection est aussi réalisable avec des drones, comme le propose **EyeSee** (2012, USA), leur système étant en production dans des entrepôts dans le groupe L'Oréal en France (vidéo).

- L'**inspection de linéaires** dans le retail avec **Bossa Nova Robotics** (2005, USA, $69,5M) avec son robot qui aligne des kilomètres dans les rayons des magasins pour analyser la présence des produits dans les linéaire et permettre le réassortiment rapide. Par contre, il ne remplit pas encore les rayons. Chaque chose en son temps ! Il serait en test dans une cinquantaine de magasins Wallmart depuis fin 2017 (vidéo) et n'aurait pas encore renversé d'enfant en bas âge ce qui est plutôt bon signe. Dans sa prochaine version, il pourrait surveiller les clients. Ou pas.

- L'optimisation des flux dans les **transits aériens** qui représentent 35% de la valeur du commerce mondial pour 1% du volume.
- Le suivi et la surveillance des **conteneurs maritimes** et des wagons dans le rail avec **Traxens** (2012, France, 1,5M€).
- Le suivi des livraisons et la prévision des délais avec **Shippeo** (2014, France, $32M) qui propose aux entreprises un suivi des processus de livraisons multi-modales alimenté par du machine learning. Celui-ci permet d'estimer les délais de livraison avec précision. La solution fonctionne évidemment grâce à la collecte de centaines de sources de données provenant des sociétés de livraison et de logistique. Leur offre logicielle est aussi proposée en marque blanche aux livreurs pour les livreurs ainsi que pour alimenter leurs services en ligne d'information pour leurs clients.

[1424] Voir 6 Applications of Artificial Intelligence for your Supply Chain., octobre 2017.

[1425] Voir L'IA fait son chemin dans la supply chain par Joe McKendrick, janvier 2020 qui cite la startup Aera Technology et dont le CEO est un français de la Silicon Valley, Frederic Laluyaux. Aera est utilisé par Unilever, Merck et ExxonMobil pour accélérer et automatiser les décisions les plus fréquentes des planners (prévisions, évitement de ruptures de stocks, ajustement des stocks, gestion des expéditions en retard).

[1426] Ces exemples sont en partie extraits de Artificial Intelligence in Logistics, un rapport commun de DHL et IBM, 2018 (45 pages).

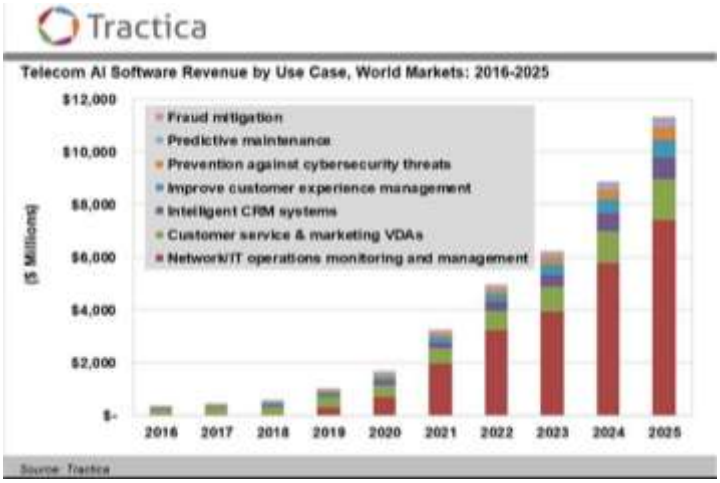
- Dans l'automatisation des **processus de dédouanement** avec des solutions juridiques développées avec **IBM Watson** et du traitement du langage.
- L'**optimisation de trajets de livraisons** proposée notamment par **LogiNext** (2014, USA, $10,6M). L'anglais **Deliveroo** utilise sa solution Frank qui est censée avoir réduit les délais de livraison de 20% au Royaume Uni à partir de 10 000 restaurants et pour 15 000 livreurs. Les livraisons sont souvent réalisées par des sous-traitants spécialisés. Ils pourront peut-être un jour mettre en route la camionnette à propulsion électrique robotisée de **Mercedez-Benz** dont l'intérieur comprend un robot de manipulation de colis qui les transmet à deux drones via des ouvertures sur le toit. Le dernier kilomètre par les airs ! Mais ce n'était probablement qu'un prototype pas forcément destiné à être commercialisé !

  Les drones viennent de **Matternet** dans lequel Mercedes a investi 562M€[1427]. Et Amazon est le premier client en vue pour ce genre de solution. C'était un prototype présenté au CES 2017 mais il ne semble pas encore opérationnel. En attendant, la livraison de proximité par drones est expérimentée par Amazon (USA, UK) ainsi que chez La Poste en France, en Isère (source).

Télécoms

Le marché des télécoms est l'un des premiers à être concerné par l'usage de l'IA dans son exploitation comme dans ses offres et leur marketing. Et pour cause, les opérateurs ont à gérer des infrastructures complexes et distribuées. Ils génèrent aussi de gros volumes de données avec l'activité des réseaux et de leurs utilisateurs. L'offre d'IA y est cependant moins lisible que sur d'autres marchés.

Elle est concentrée sur quelques acteurs à commencer par les grands équipementiers (**Huawei**, **Cisco**, **Nokia**, **Ericsson**, **ZTE**), puis diverses startups et des prestataires de services et enfin, les opérateurs télécoms eux-mêmes, qui développent souvent leurs propres solutions sur mesure ou conjointement avec leurs équipementiers et prestataires de services[1428]. Ils jouent souvent le rôle d'intégrateur. D'autres comme **Free** ont tendance à exploiter des logiciels open source et à développer les leurs.



Parmi les éditeurs de solutions logicielles à base d'IA destinées aux opérateurs télécoms, domineront essentiellement ceux qui fournissent des outils de gestion des opérations réseau et IT, suivis par ceux qui proposent des outils liés à la relation clients[1429].

Comme partout, l'IA est une grosse boîte à outils exploitée application par application, et de manière disparate. Il n'y a pas encore de Gosplan ou de méta-architecture d'IA pour gérer l'ensemble des systèmes d'information des opérateurs télécoms.

Les principaux éditeurs de logiciels des télécoms sont soit génériques soit spécifiques à ce marché.

[1427] Nombreux details ici : http://www.businessinsider.fr/us/mercedes-electric-vision-van-drone-delivery-service-photos-2017-3/.

[1428] Comme Nokia qui est partenaire de China Telecom dans l'IA. Voir Nokia and China Mobile to set up joint AI*5G lab for further research using artificial intelligence and machine learning in 5G networks, juillet 2018.

[1429] Source : étude marché de Tractica, 2018, vue dans Report says AI Sales in Telecom to Reach $36,7B By 2025, avril 2018.

Avec notamment **Afiniti** (association agents/clients en centres d'appels), **AIBrain** (IA générique de raisonnement, moteurs de règles et réseaux d'agents), **Anodot** (détection d'anomalies en machine learning), **Aria Networks** (optimisation de réseaux), **B. Yond** (et sa plateforme d'analytics et de supervision de réseaux Infinity), **Cardinality** (analytics engines), **Guavus** (analytics temps réel), **Intent HQ**, **Ipsoft** (chatbot d'assistance d'agents de helpdesk), **Skymind** (déploiement distribué de solutions d'IA), **Subtonomy** (monitoring de réseau mobile), **Tupl** (monitoring de réseaux), **Sysmech** (outils de gestion de réseaux) et **Wise Athena** (outil de pricing).

Réseau et exploitation

Le déclencheur d'un usage intensif de l'IA chez les opérateurs sera probablement le déploiement et l'exploitation de réseaux sans fil 5G dans les douze ans qui viennent. Le niveau de complexité de ces réseaux augmente d'un ordre de grandeur par rapport à la 4G.

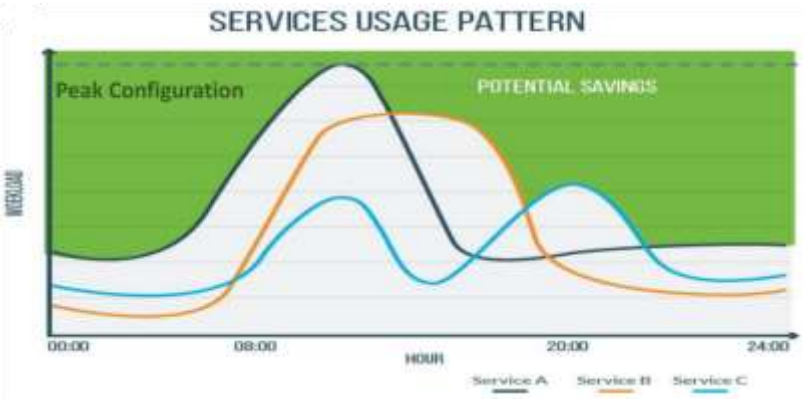
La 5G va voir se généraliser des concepts tels que le **SDN** (Software Defined Network), les **SON** (Self-Organized Networks), et la **NFV** (Network Function Virtualization)[1430], pour décrire des architectures de pilotage de réseaux reconfigurables qui reposent sur des serveurs génériques et sur une gestion très dynamique des réseaux et de leurs infrastructures, avec pour finalité, une reconfiguration dynamique des réseaux en fonction du trafic[1431].

Les réseaux 5G vont nécessiter de prédire l'évolution du trafic avec une forte granularité pour allouer les fréquences dans la 5G notamment dans les zones très denses, et pour maximiser l'usage de nombreuses fréquences traditionnelles (en-dessous de 5 GHz), pré-millimétriques (5 à 28 GHz) et millimétriques (28 GHz et au-delà).

D'autres enjeux sont à prendre en compte comme la réduction de la consommation d'énergie des serveurs hors des heures de pointe d'usage du réseau. Tout cela en temps réel, au gré de l'évolution de la demande.

Les réseaux vont aussi voir augmenter la part des vidéos qu'ils transportent, avec des exigences non négligeables en termes de qualité de service qui va nécessiter encore plus de coordination entre les techniques d'adaptive streaming et les liens avec les CDN (Akamai and co).

L'IA peut intervenir pour prédire les évolutions de la demande de télécommunications et aider les opérateurs à préparer les infrastructures à s'y adapter, de plus en plus automatiquement. Cela nécessitera d'ailleurs probablement des techniques d'apprentissage par renforcement pour affiner les modèles prédictifs au fil de l'eau. C'est une approche retenue par **China Telecom**[1432] (*ci-contre*).



[1430] Et cela comprend aussi ONO (Online Network Optimization) qui est proposé dans Model-Driven Artificial Intelligence for Online Network Optimization 2018, (10 pages) et qui consiste à configurer dynamiquement les réseaux avec du deep learning alimenté par les données des SDN.

[1431] Voir aussi Defining closed-loop AI mechanisms for network management de l'ETSI, 2018 (55 slides) ainsi que Artificial Intelligence for 5G: Challenges and Opportunies de Merouane Debbah du centre de recherche de Huawei en France, 2018 (64 slides). Enfin, les usages du machine learning dans la gestion des réseaux sont bien décrits dans Deep Learning in Mobile and Wireless Networking, a survey, par Chaoyun Zhang, Paul Patras et Hamed Haddadi, janvier 2019 (67 pages).

[1432] AI and other emerging ICT technologies bring new development opportunities for telecom operators de Xiaoou Liu, China Telecom, avril 2018 (32 slides).

En amont du pilotage, l'IA – et même un jour, le calcul quantique – peut intervenir pour optimiser le placement des antennes et minimiser leur nombre tout en maximisant leur couverture, le problème mathématique sous-jacent étant celui de la résolution du théorème des quatre couleurs[1433].

La supervision des réseaux des opérateurs va utiliser de plus en plus de machine learning. Le Coréen **SK Telecom** utilise d'ailleurs depuis 2017 une telle solution pour gérer ses opérations réseaux et améliorer la qualité de service et l'état de ses infrastructures, dénommée Tango (vidéo).

Ericsson lançait en 2020 des solutions de gestion de la consommation énergétique des infrastructures réseaux à base de machine learning, Energy Infrastructure Operations, applicables en particulier aux infrastructures de la 5G. L'objectif est bien entendu de faire des économies d'énergie.

Le machine learning peut aussi servir à détecter les contenus piratés, notamment ceux qui sont diffusés avec des protocoles de streaming vidéo[1434].

Enfin la maintenance des infrastructures peut aussi faire appel à du machine learning pour identifier les infrastructures à maintenir ou remplacer avant que des pannes surviennent.

Sysmech (1994, UK) propose ainsi des outils de supervision de réseau à base de machine learning qui peuvent détecter les émetteurs du réseau qui sont en panne pour les redémarrer.

Cela passe même par l'inspection visuelle des antennes par des drones comme chez **Aerialtronics** (2012, Pays-Bas, 3M€), une technique qui est notamment utilisée par **AT&T** aux USA (vidéo) depuis 2016.

Offres marchés

Le second domaine où l'IA intervient chez les opérateurs télécoms est dans leurs offres commerciales, surtout grand public.

Du côté du fixe, les outils de **recommandation de vidéos** s'appuient souvent sur du machine learning et sont légion, même s'ils sont souvent issus d'éditeurs de logiciels indépendants.

Les set-top-boxes sont aussi commandables à la voix. Par contre, les opérateurs interviennent moins sur les mobiles eux-mêmes du fait de la maîtrise des outils logiciels par les plateformes Android et iOS.

Au-delà de la recommandation de contenus divers, l'IA peut intervenir à plusieurs endroits de la chaîne produit, que ce soit pour de l'upsell ou du crosssell, dans les boutiques sur le terrain et dans les applications de web-to-store. Ce sont des usages génériques du retail applicables au métier des opérateurs télécoms.

Les opérateurs mettent aussi en place des solutions à base d'IA pour la **détection de différentes formes de fraude**, surtout aux USA. Cela concerne l'usurpation d'identité avec de fausses pièces d'identités utilisées pour souscrire des abonnements.

Chez **Deutsche Telekom**, on utilise depuis 2018 les logiciels biométriques de l'Américain **Nuance** pour réaliser cette authentification des utilisateurs lorsqu'ils appellent le support en ligne. La solution permet aussi de faire sa demande en langage naturel pour être orienté vers la bonne personne, au lieu de passer par une interminable suite de questions et de numéros à saisir. Cette solution de Nuance est également déployée en Turquie par **Vodafone**.

Mais l'opérateur allemand est aussi partenaire avec **Orange** pour la conception de l'agent vocal Djingo et ses incarnations physiques comme le Halo Magenta Speaker de l'Allemand et le Djingo

---

[1433] Voir Apprentissage machine pour l'optimisation énergétique des réseaux cellulaires hétérogènes sans-fil : une approche bandit à bras multiples par Navikkumar Modi & AL, 2019. Le papier décrit une méthode de machine learning permettant d'optimiser la consommation d'énergie des antennes relais.

[1434] Voir Machine learning working against video piracy, mars 2020.

Speaker d'Orange. Il devait piloter différents objets connectés ainsi leur offre Orange Bank. Les deux opérateurs partagent à 50/50 la R&D. Le Smart Speaker Djingo était annoncé à 49€. La mise en production de l'agent vocal a subit des retards. Présenté pour la première fois en 2017, il était effectivement lancé deux ans plus tard, mi novembre 2019, pour être abandonné en octobre 2020.

Avec la 5G, les opérateurs télécoms sont tentés de mettre le nez dans des marchés verticaux qui seront très consommateurs de briques logicielles d'IA : avec les véhicules connectés et autonomes, dans la santé et la ville intelligente. Les opérateurs sont impliqués dans ces domaines comme des fournisseurs d'infrastructures. Mais parfois, ils s'engagent avec du service et de l'intégration comme chez Orange Business Services ou chez Bouygues Telecom via leur filiale Objenious.

Relation client

On pense évidemment en premier aux chatbots qui sont maintenant très courants chez les opérateurs télécoms. On en trouve chez **Bouygues Telecom**, chez **Orange**, chez **Telefonica** et **Vodagone** (avec Tobi[1435]). Ils permettent notamment d'accéder aux bases de connaissance du support technique.

Aux USA, on en trouve chez **Spectrum**, l'opérateur issu de la fusion de Charter Communications et Time Warner Cable, propose l'assistant virtuel Ask qui gère le support commercial et technique.

Le câblo-opérateur **Century** utilise depuis 2016 son agent Angie développé avec **Conversica** (2007, USA, $72M) et qui analyse 30 000 mails entrants par mois.

Tous ces outils couplés aux indicateurs d'utilisation des services télécoms fixes et/ou mobiles permettent de faire du machine learning pour identifier les clients susceptibles de quitter le service (le « churn »). A charge ensuite de 1) comprendre la situation exacte du client et 2) lui proposer une éventuelle offre adaptée différente du package dont il dispose.

L'autre approche consiste à associer chatbots et agents humains dans les centres d'appels, ce que propose **Afiniti** (2006, USA, $137M). Ils utilisent un modèle permettant l'appariement des clients avec les bons agents dans les centres d'appels. La solution à base de machine learning exploite l'historique client. On peut se demander si ce genre de solution n'est pas difficile à déployer en l'état dans l'Union Européenne du fait du RGPD. Mais il y ont des clients, y compris en France.

En dernier lieu, citons l'usage des robots **Pepper** pour l'accueil dans les boutiques de Softbank au Japon, qui y sont déployés depuis 2014 (*ci-contre*, dans celle d'Otomesando à Tokyo en octobre 2014). Quel retour en avons-nous ? Le déploiement était explicable par le fait que l'équipe conceptrice de Pepper issue de l'acquisition du Français Aldebaran Robotics fait aussi partie du groupe Softbank. Mais les scénarios d'usage de Pepper en boutique restent limités. Qu'en reste-t-il une fois passé l'effet de surprise et de découverte initiale ?

Médias et contenus

Les médias font partie de ces métiers qui ont été particulièrement bousculés par l'irruption des outils numériques, d'Internet et des médias sociaux. Leur chiffre d'affaire a baissé, leurs revenus publicitaires ont en parti migré vers d'autres acteurs, que ce soit les GAFA, les services en ligne d'offres d'emplois ou dans l'immobilier.

[1435] Voir http://labs.vodafone.co.uk/case-studies/tobi et leur vidéo.

Les moyens baissant, ceux qui sont alloués aux journalistes pour mener leurs enquêtes ont fondu d'autant. Nombre de médias ont décliné, surtout dans la presse quotidienne nationale et régionale.

L'adoption de nouvelles méthodes de travail ne s'est pas faite sans mal. Les rédactions digitales étant trop souvent séparées des rédactions historiques. Les premières ont adopté des méthodes peu recommandables, republiant des informations sans prendre le temps d'enquêter, générant des effets de caisse de résonnance à ce que l'on appelle maintenant les vraies fausses nouvelles. Les seconds ont de leur côté ignoré les outils permettant à leurs écrits d'être mieux diffusés.

Sur ce arrive la vague de l'IA qui entraîne tout sur son passage et qui peut à son tour bouleverser une fois de plus les médias[1436]. Avec la crainte qu'elle génère une nouvelle vague déflationniste du côté de l'emploi.

Nous allons donc voir ici comment les médias, et surtout la presse, peuvent tirer parti de l'IA à la fois pour la production de contenus, pour leur diffusion et pour leur monétisation[1437] et si cela prête à conséquence.

L'IA est comme Internet et Google Search. Elle peut entraîner le meilleur comme le pire. Elle permet aussi bien d'améliorer la qualité de ses contenus que de se désengager de ce point de vue-là. Elle peut aussi véhiculer des biais. Comme elle est entraînée par les contenus existants et leur appréciation par le public, elle peut enfermer les producteurs et les consommateurs dans des bulles de contenus, comme le font déjà les réseaux sociaux. Il faut donc se garder de prêter plus d'intelligence et de créativité à l'IA quelle n'en a.

production	diffusion	monétisation
détection de signaux faibles	indexation, recherche et curation de contenus	ciblage d'audiences
vérification des sources	analyse d'impact émotionnel des contenus et parcours	recommandation contenus
rédaction de dépêches	modération commentaires	optimisation de sites web
résumés automatiques	analyse de compétitions sportives	analyse d'impact publicitaire
cadrage et montage vidéo	formatage de contenus	optimisation référencement
sous-titrage	coloriage, résolution	ajustement du freemium
accompagnement musical	suivi émotionnel	
habillage jeux vidéo	détection de fake news	
big data d'enquêtes complexes		
présentateurs virtuels		
fake videos and pictures		

L'IA fournit surtout une boîte à outils qui peut potentiellement faire gagner du temps[1438]. Elle fournit une aide aux journalistes pour analyser les données et détecter des tendances à partir de sources d'informations multiples allant des sources ouvertes habituelles aux sources inédites comme les données publiées par Wikileaks[1439].

[1436] Voir IA dans les médias, un peu, beaucoup, passionnément ? Une cartographie des applications d'Intelligence Artificielle par Kati Bremme, 2019 et Médias automatiques ou complicité hommes-machines ?, par Eric Scherer, mai 2019.

[1437] Voir Et si les médias redevenaient intelligents ?, mai 2017. Et aussi cette étude Artificial Intelligence in the Media and Creative Industries Position Paper, 2018 (30 pages).

[1438] Voir Journalisme augmenté à l'IA, un état des lieux par Kati Bremme, novembre 2019.

[1439] Voir Comment le machine learning et la data science confèrent à Bloomberg un avantage concurrentiel par Mark Samuels, novembre 2018.

Elle peut aussi aider à convertir les données en texte, les textes en contenus audio et vidéo. Elle permet d'analyser des objets, des images, de reconnaître des personnes dans des photos et les vidéos, ce qui est potentiellement utile pour comprendre des situations captées par des photojournalistes.

Nous évoquerons aussi le rôle de l'IA dans d'autres types de contenus, en premier lieu la musique et les jeux vidéo. L'IA apporte de nouveaux outils de génération de contenus qui font peur à certains. En fait, il s'agit d'une extension de la palette des outils mis à la disposition des créatifs et à charge pour eux de s'en emparer et de rester créatifs comme l'ont été ceux qui ont adopté en leur temps les outils de dessin vectoriel type Illustrator ou de retouche photo type Photoshop[1440].

Recherche

Les nombreux moteurs de recherche documentaires génériques du marché peuvent être exploités par les journalistes s'ils peuvent s'en équiper. Il en existe quelques-uns qui sont dédiés aux journalistes, comme Salient de **Lore.ai** (2016, France/USA), un outil d'analyse de contenus qui sert à réaliser des recherches, des liens entre documents, de les classifier et de les exploiter, le tout sur plusieurs sources d'information et plusieurs langues.

Newsbridge (2016, France, 1,5M€) réalise de l'indexation automatique et temps réel de vidéos par analyse des voix qu'elles contiennent avec du speech-to-text et des images.

Cela permet ensuite de faire facilement des recherches dans ses rushes de reportages ou dans les vidéos déjà publiées (exemple de recherche *ci-contre*).

Écrit

Nombre de startups sont apparues qui automatisent la production de contenus. Comme presque partout dans l'IA, ce qu'elle produit ne vient pas de nulle part mais réutilise des contenus existants créés par de vrais gens. Et elle les assemble en observant la manière dont les contenus existants sont assemblés.

Nous avons pu balayer quelques startups spécialisées dans la génération automatique de textes et de résumés dans la rubrique sur le traitement du langage. Une partie d'entre elles ciblent les médias, pour ce que l'on appelle le « robot journalism » comme le **Syllabs** (2006, France, 2M€) et sa solution Syllabs Media utilisée chez Radio France, à l'AFP ou dans le journal Le Monde, ou encore **Textomatic** (2010, Allemagne, $40M). Syllabs Media est voisin de la solution d'Yseop et génère des textes à partir de données structurées et chiffrées. Le journaliste doit ensuite modifier le texte pour lui ajouter de l'émotion et du vivant. Syllabs produit notamment des contenus multilingues pour les médias qui exploitent les résultats d'élections, en France dans un premier temps, des contenus touristiques locaux ainsi que des annonces immobilières.

Très souvent, les startups qui ciblent initialement le marché des médias s'en détournent progressivement ou partiellement parce que ce marché est moins solvable et dynamique que ceux de la finance ou du marketing.

Les robots rédacteurs ne font pas de véritable journalisme. Ils génèrent des textes répétitifs sur de gros volumes de données comme pour produire les résultats d'élections à l'échelle locale, dans le cadre de compétitions sportives ou pour la météo.

[1440] Voir L'intelligence artificielle créative…vers de nouveaux horizons d'Olivier Reynaud, septembre 2018.

Ils transforment le plus souvent des données numériques en phrases avec des modèles plus ou moins flexibles. Mais les générateurs de langage peuvent de plus en plus tenir compte du contexte des données pour choisir les bonnes formulations.

C'est le cas avec la publication en mai 2017 d'un article du Los Angeles Times annonçant un tremblement de terre d'échelle 3.8 (vidéo)[1441].


Los Angeles Times

3.8 earthquake shakes Los Angeles area

Staff writer Shelby Grad contributed to this report.

L'article n'a pas été écrit par un système à base d'IA mais par un petit programme dénommé **ClickBot** développé par un journaliste du LA Times.

Un journaliste a ensuite complété l'article à la mano. C'est l'émergence du journaliste assisté par l'IA[1442]. Mais cela peut aller plus loin, notamment pour les journalistes qui font de la curation. Microsoft Corp n'a ainsi pas hésité à se séparer en 2020 de journalistes qui en faisaient pour MSN, pour les remplacer par des robots logiciels, non sans mal car cela peut générer parfois quelques bévues[1443].

L'agence **Associated Press** publie depuis 2015 des dépêches créées par des robots journalistes pour les annonces standardisées, notamment dans l'actualité financière (vidéo). Le **New York Times** exploite de son côté l'outil **Editor** depuis 2015 qui permet aux journalistes d'insérer des tags dans les articles, qui vont ensuite servir à extraire l'information associée.

Les nouveaux systèmes de génération de texte utilisant des moteurs tels que GPT-3, sorti mi 2020, ont probablement des fonctionnalités plus avancées qui feront leur apparition dans les outils destinés aux journalistes. Les journalistes des médias écrits ont aussi besoin d'exploiter les médias chauds (radio, TV). La simple visionnage de vidéos est trop long pour une rédaction dans la presse écrite. Les outils de transcript de vidéos en texte sont donc les bienvenus. C'est une fonction standard dans **YouTube** (*ci-dessous*) !

[1441] L'article en question : 3.8 earthquake shakes Los Angeles area par Quakebot, mai 2015.

[1442] Voir The Rise of the Robot Reporter par Jaclyn Peiser, février 2019, Will AI Save Journalism - or Kill It?, avril 2019 et IA et médias : une innovation plus qu'une révolution par Jean-Dominique Seval, mars 2019.

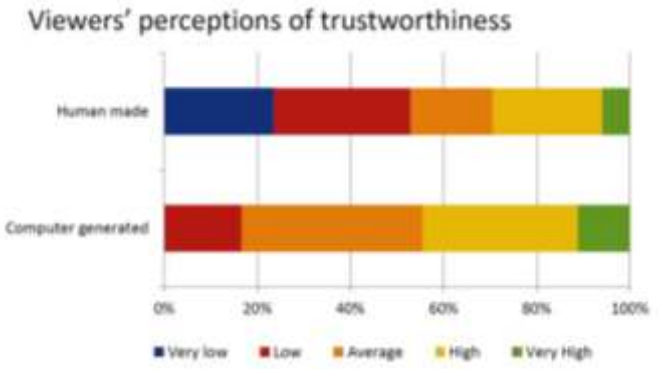
[1443] Voir Microsoft sacks journalists to replace them with robots par Jim Waterson, mai 2020.

D'autres solutions de génération de contenus dédiés aux médias ont vu le jour sur d'autres types de contenus. **Valossa** (2015, Finlande, $2,7M) propose ainsi une solution en cloud de reconnaissance d'images dans les vidéos adapté aux besoins des broadcasters. Elle permet l'interprétation de vidéos, détecte les personnes, leur verbatim et les thèmes couverts (*ci-dessous à droite*). Elle ajoute des métadonnées aux scènes analysées exploitables dans les outils d'analytics voire pour les générateurs de guides de programmes.

Lobster (2013, UK, $2M) fait de la curation de contenus issus des médias sociaux pour alimenter les agences de communication et entreprises.

Le breton **Mediawen** (2014, France) gère la traduction de contenus vidéo en temps réel en s'appuyant sur IBM Watson puis text to speech, en voix de synthèse ou sous titrage.

Une étude américaine montre que les lecteurs font plus confiance à des articles produits par des robots que par des journalistes[1444]. Cela mérite évidemment un peu de recul car les articles produits par les robots journalistes ne font le plus souvent que transformer des données chiffrées en phrases et ne véhiculent donc pas d'opinion ou de jugement de valeur. Sauf … si les données sont fausses ! Et, sans surprise, les journalistes n'aiment pas les robots journalistes[1445] !



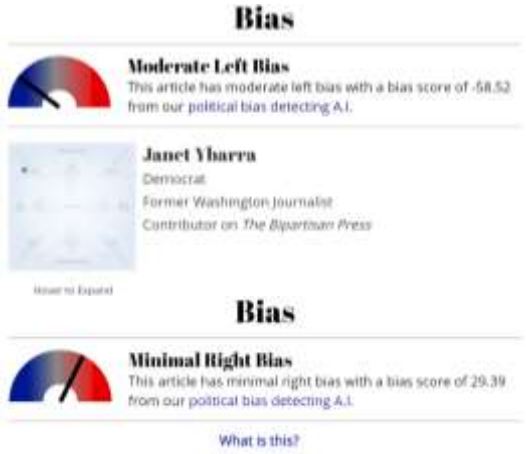
Knowhere (2015, USA, $2M) est une startup qui résout le problème en sélectionnant des articles structurellement neutres. C'est une sorte d'outil de curation qui jauge de la véracité des articles. Il sélectionne d'abord les sujets en fonction de leur popularité, ce qui est déjà un biais humain soit dit en passant, ne retient que les sujets cités dans au moins cinq sources fiables, en éliminant les sites conspirationnistes. Un article synthétique est ensuite automatiquement réalisé par mash-up des articles sélectionnés en retenant les faits et en éliminant les avis. Un journaliste intervient en bout de course pour valider l'ensemble. Notez cependant que la source de tout cela est faite d'articles écrits par des journalistes en chair et en os.

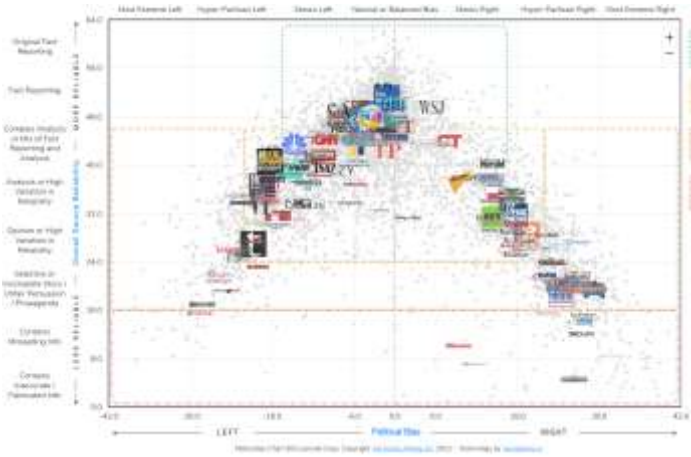
Bipartisan Press (site) est un site d'information américain lancé en 2018 qui a développé une brique d'IA qui classe le biais politique de ses articles sur deux dimensions : l'orientation gauche/droite et l'intensité du biais ("minimum" à "extrême"), avec une précision de 96% et un écart type de 7%. Il propose ses « Bias Analyzer API » à des sites tiers et prévoit de publier une extension pour Chrome permettant l'analyse de n'importe quelle page web. Cela ne va pas pour autant dépolariser les débats, surtout aux USA. Si vous allez sur FoxNews, vous savez à l'avance que c'est biaisé. Pareil sur le Washington Post ou sur DailyKos (un blog démocrate).

Un graphe positionne d'ailleurs les médias en ligne et TV américains sur cette échelle de biais, de l'ONG américaine **Ad Fontes Media**. Il illustre le fait que plus les sources d'informations sont biaisées, moins elles sont fiables.

[1444] Source du schéma à droite : The Artefacts of Automated Journalism: Producers' Perspectives and Audience Assessments de Neil Thurman.

[1445] Voir Human journalists hate robot journalists, says new report, mars 2017, qui propose au passage un petit test de détection de brèves rédigées par des robots et des journalistes. Il n'est pas trop difficile d'obtenir 5 sur 5 au test. Ce qui est rassurant pour les journalistes !





Des chercheurs de **Thomson Reuters** et **Alibaba** génèrent de leur côté des news inédites (« scoops ») à partir des flux détectés dans Twitter[1446] dans l'application Reuters Tracer. Le système repère les nouvelles, les classifie, les annote, les mets dans l'ordre, rédige un article et le publie. Et il le fait plus vite que les agences de presse traditionnelle. Et même s'il faut peut-être éditer le résultat, le processus sera toujours plus rapide.

Une approche voisine a été lancée en 2020 avec le site **NotRealNews** qui comprend de courts articles générés par une IA dénommée Wordflow AI. Le propos est de montrer qu'un tel système peut aider les journalistes à gagner du temps dans la rédaction d'articles. Pas forcément à créer des fake news [1447]!

BingeBooks (2020, USA) propose de la recommandation de livres à binger comme des séries TV. Bref, pour les lire. Il est doté d'une IA dénommée Marlowe et développée par **Authors AI** (2020) à base d'analyse du langage qui réalise un profiling des textes des ouvrages et réalise des critiques d'une vingtaine de pages automatiques. Marlowe est censé analyser les traits de caractère des personnages, les arcs narratifs, le rythme et la ponctuation du texte. Bingebooks a la particularité d'avoir été lancé collectivement par plus d'une centaine d'auteurs américains.

Enfin, citons ces IA qui génèrent des nouvelles ou des poèmes[1448]. Nombre de commentateurs s'en émerveillent alors qu'il ne s'agit que de systèmes probabilistes et combinatoires qui n'ont aucune intelligence symbolique ni capacité émotionnelle.

Le fruit de la combinatoire et d'une partie de hasard suffit à générer des créations et à faire illusion. On oublie évidemment que l'Homme fait le tri en sortie de machine pour ne conserver que cela qui a l'air exploitable et publiable. Les contenus poubellisés sont rarement présentés dans ces productions [1449]!

Audio et musique

L'industrie musicale tire aussi parti de l'IA pour la production et la diffusion de contenus. C'est une grande consommatrice de réseaux génératifs. Les outils de création de musique artificielle entraînés par l'ingestion de wagons de répertoires musicaux sont légion.

[1446] Voir Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data, novembre 2017 (11 pages).

[1447] Voir This Site Uses AI to Generate Fake News Articles par Dan Robitzski, février 2020.

[1448] Voir par exemple When an AI Goes Full Jack Kerouac - A computer has written a "novel" narrating its own cross-country road trip, de Brian Merchant dans The Atlantic, octobre 2018. Et puis aussi Turing Test Passed : Using Computer Generated Poetry, janvier 2015. L'auteur ne sait visiblement pas ce qu'est le test de Turing ! C'est un test portant sur un dialogue avec une machine qui serait impossible à distinguer d'un homme. Ici, il ne s'agit pas de dialogue. Voir aussi le générateur de poèmes de Ray Kurzweil : Ray Kurzweil's Cybernetic Poet: How it works.

[1449] Voir Comment l'intelligence artificielle peut-elle aider la création par Benjamin Hoguet, 2019.

Pour l'instant, ils ont bien du mal à créer de la musique de qualité mais la discipline fait des progrès. C'en est encore très souvent au stade de la musique d'ascenseur ou de jingles publicitaires. Mais cela progresse doucement.

Les prouesses de ces IA sont d'ailleurs souvent exagérées. En effet, ce sont des IA pilotées par des humains qui restent juges du résultat. On se garde bien de préciser qu'une prouesse est la conséquence de la poubellisation de nombreux tests non concluants. Il y a donc bien une oreille humaine qui fait le tri. L'ADN du hit n'a pas encore été trouvée et laisse encore de la place aux véritables artistes. Et lorsqu'il n'y a pas de tri, le résultat est très moyen, comme pour ce générateur de solo de basse.

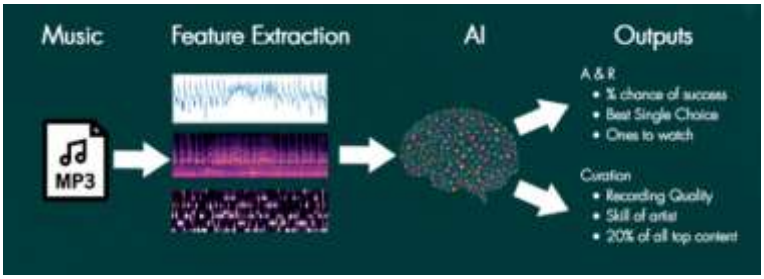
Voici un inventaire de quelques unes des startups ou projets de recherche de ce domaine florissant :

- **Amper Music** (2014, USA, $9,1M) est un site en ligne qui compose automatiquement de la musique via de l'IA (vidéo). Il faut tout de même le paramétrer pour indiquer ses souhaits en termes d'instruments, de tempo, de style et de durée. C'est pratique pour créer des compositions qui vont alimenter les vidéos de démonstration de startups et éviter les habituelles musiques d'ukulélé qui les accompagnent régulièrement. La startup a plusieurs concurrents directs comme **Jukedeck** (2012, UK, $3,4M, exemple, acquis par TikTok en 2019), **Melomics Media** (2012, USA, exemple), **Aiva Technologies** (2016, Luxembourg, 2,2M€) (exemple 1 de musique symphonique et exemple 2 de musique rock), **Melodrive** (2017, Allemagne, exemple).
- **DeepComposer** est un système de création d'accompagnement musical à base de GAN lancé par Amazon en décembre 2019. Il est fourni avec un clavier MIDI assez classique en apparence à deux octaves qui se connecte en USB à votre micro-ordinateur (vidéo). L'application associée permet d'enregistrer une simple mélodie et ensuite, de la transformer en œuvre polyinstrumentale avec un réseau de neurones génératif exécuté dans le cloud AWS avec le choix du style de musique (rock, pop, jazz, classique)[1450]. L'environnement de développement d'AWS SageMaker permet alors de créer son modèle génératif personnalisé. Le résultat est assez approximatif. Disons que c'est même de la soupe !
- **Landr** (2012, Canada, $345,3M) propose un service en cloud d'automatisation du mixage audio, qui va créer des morceaux de musique agréables à l'écoute (vidéo).
- **Popgun.ai** (2017, Australie) utilise le deep learning pour apprendre les bonnes règles musicales à partir de compositions humaines et pour enrichir des compositions existantes. La démonstration de leur prototype Alice est sympathique (vidéo) mais pas forcément facile à mettre en pratique.
- **Pacemaker** (2011, Suède, $4,5M) est un DJ à base d'AI qui exploite les contenus de services de musique en ligne comme Spotify. Il s'agit ici de gérer un bon séquencement de morceaux. L'entraînement peut avoir été réalisé avec des séquences existantes. On ne doit pas être trop loin du concept des réseaux de neurone à mémoire de type LSTM.
- **MXX** (2015, UK) propose une solution logicielle qui édite une musique automatiquement pour la synchroniser avec une vidéo faisant ainsi gagner beaucoup de temps au monteur.
- **Muzeek** (2015, USA, $500K) génère directement une musique originale et la synchronise avec une vidéo. Le marché visé est surtout celui de la création de spots publicitaires.

[1450] Voir AWS announces DeepComposer, a machine learning keyboard for developers, décembre 2019.

- **Google** a créé son propre générateur de sons à base de deep learning, Nsynth (Neuron Synthetizer), issu du projet Magenta. Il est couplé à un instrument de pilotage, Nsynth Super (vidéo) qui pourrait un jour faire partie de la panoplie des compositeurs et DJs. Mais ce n'est pas encore un produit. C'est juste un projet de chercheurs[1451].

- **Musiio** (2018, Singapour, $1M) a développé un outil qui permet aux labels d'édition d'artistes de sélectionner les artistes, les morceaux à éditer et les singles à mettre en avant. Il s'entraîne avec l'ingestion de dizaines de milliers de musiques avec du deep learning.



  Cela rappelle le projet du chercheur de la BBC Armand Leroi publié en 2017 et ayant donné lieu à un documentaire de la chaîne[1452]. Le projet était "inconclusif", à savoir que si le machine learning avait bien permis de déterminer quelques caractéristiques des morceaux qui plaisaient selon les époques, il n'avait pas permis de créer pour autant un nouveau hit.

- **Huawei** faisait terminer la Symphonie Inachevée de Schubert (numéro 8 en si mineur) par une IA et présentait le résultat en février 2019 à Londres. Le tout était semble-t-il réalisé avec un smartphone (vidéo). Un projet voisin était lancé fin 2019 par Deutsche Telekom pour terminer la 10e symphonie de Beethoven à partir des quelques notes et croquis laissés par le compositeur avant sa mort. La symphonie devait être révélée le 28 avril 2020 à l'occasion du 250e anniversaire de sa naissance. On ne saura pas ce qu'il en est car l'événement n'a visiblement pas eu lieu à cause du confinement covid-19 qui sévissait à ce moment-là. Faudra-t-il attendre le 260e anniversaire pour que l'IA soit mise au point ? Les prémices ne semblaient en tout cas pas prometteuses[1453].
- **Nami Lab** (2016, Italy, 250K€) a développé l'application Yalp qui fait de la transcription en partition commentée de toute musique. Cela peut aider les musiciens, notamment les guitaristes, à reprendre des morceaux de répertoire.
- Le laboratoire **Sony CSL** situé à Paris a créé Flow Machines, une IA génératrice de pop-music. En 2016, elle créait Daddy's Car, un succédané des Beatles sauce Georges Martin s'étant emmêlé les paluches dans la table de mixage d'Abbey Road ou Phil Spector ayant mélangé un peu trop de drogues douces (vidéo)[1454]. Plus récemment, elle composait l'album « Hello World » (vidéo) mais avec l'aide d'artistes.

[1451] Voir Artificial intelligence can now emulate human behaviors – soon it will be dangerously good par Ana Santos Rutschman, avril 2019, à propos d'une IA de Google qui génère de la musique en s'entraînant avec le catalogue de Bach.

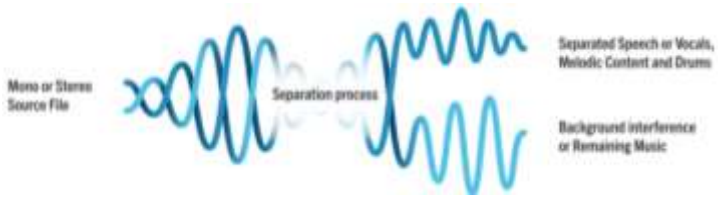
[1452] Voir The Secret Science of Pop, 2017.

[1453] Voir La 10e symphonie de Beethoven sera composée par l'intelligence artificielle par France Info, décembre 2019.

[1454] L'histoire évite d'indiquer combien de morceaux créés par l'IA en question ont du être poubellisés avant la production de celui qui a été retenu ! In fine, l'Homme conserve ce qui lui semble être pertinent. L'IA n'a pas encore cette capacité de jugement artistique et émotionnelle. Ces IA créatrices doivent être surtout percues comme de nouvelles palettes pour les compositeurs. Les IA créatrices fonctionnent encore par apprentissage par renforcement, ce dernier provenant des créateurs humains.

- Toujours chez **Sony**, une IA ajoute la batterie à un morceau de musique existant. Leur système a été entraîné avec 665 chansons dans plusieurs genres avec des pistes séparées pour chaque instrument[1455].
- Le chanteur de The Voice en Norvège **Thomas Holm** créait une chanson de Noël via une IA exploitant une bibliothèque de chansons du même type, avec paroles et musique[1456]. Le groupe **YACHT** créait un album complet, Chain Tripping, en s'aidant d'une IA exploitant le projet Magenta de Google en 2019 et qui exploite pour s'entraîner leurs 82 chansons existantes[1457].
- **Warner Music** serait le premier label de musique ayant "signé" un contrat avec une IA, en fait, avec l'équipe créatrice de l'application **Endel** qui génère de la musique d'ambiance et de relaxation. L'application est accessible via Amazon Alexa, avec cinq albums produits en 2019[1458].
- En 2019, la chaîne YouTube **Relentless Doppelganger** se mettait à diffuser du heavy metal en continu généré par une IA. Il s'agit d'une chanson qui tourne en boucle et sans fin. L'IA a été développée par CJ Carr et Zack Zukowski et entraînée avec des contenus issus du groupe canadien Archspire. J'aurais préféré qu'ils testent cela avec Led Zeppelin !
- **Audionamix** (2003, France, $5,9M) utilise l'IA pour optimiser le mixage audio. Leur logiciel ADX Technology est capable d'isoler une mélodie ou un chant d'un enregistrement mixé. C'est notamment utile pour les DJ qui font des remix (vidéo). Cela peut aussi servir à remixer un contenu stéréo pour l'adapter à des systèmes multicanaux (Dolby 5.1, Dolby Atmos, etc)[1459]. C'est aussi ce que propose le projet Spleeter open source de Deezer qui utilise TensorFlow pour dissocier les différents instruments dans un morceau musical. Lancé fin 2019 et inrégrable sous forme de plugin dans certains mélangeurs audio, il sépare la partie vocale, les basses, la batterie (voir la vidéo du test réalisé par Korben) et le reste. Le résultat est plus ou moins convaincant selon les morceaux. **Izotope RX** (2001, USA, $19,5M) réalise un traitement du même genre (vidéo).

  Audionamix a été créée bien avant la grande vague de l'IA des années 2010. Ses solutions ont été adaptées pour exploiter les dernière prouesses du machine learning et du deep learning.



- Le projet **Audio Awesome** d'Adobe s'appuie sur Sensei AI pour améliorer l'extraction de voix propres dans des enregistrements audio. Le système s'appuie sur un WaveNet (DeepMind) pour le débruitage de l'audio complété par un GAN (réseau génératif)[1460].
- On peut apprécier le projet open source français **AdBlock Radio** qui sert à éviter la publicité dans l'écoute de la radio. Il a été lancé par Alexandre Storelli, un ingénieur de l'X. Il n'est évidemment pas certain que ce logiciel ait un modèle économique extraordinaire.

[1455] Voir Sony's AI drums up beats for songs par Kyle Wiggers, mai 2019 (vidéo).

[1456] Voir Artificial Intelligence Meets Christmas Spirit, 2017.

[1457] Voir How YACHT Used Machine Learning to Create Their New Album, Wired, septembre 2019.

[1458] Voir Warner Music signs first ever record deal with an algorithm, mars 2019.

[1459] Voir Artificial Intelligence in Audio Event at Dolby Soho, mai 2019 (1h42), une conférence de Dolby sur les usages de l'IA dans l'audio.

[1460] Voir Adobe's Project Awesome Audio cleans up recordings with a single click par Billy Steele, novembre 2019, qui fait référence à Perceptually-motivated environment-specific speech enhancement par Zeyu Jin & Al, 2019 (5 pages).

- Pour ce qui est du futur du karaoké, des chercheurs du **Facebook AI Research** et de l'**Université de Tel Aviv** ont créé un système qui convertit la voix d'un chanteur dans celle d'un autre chanteur.

  Le système fonctionne en mode non supervisé, mais tout de même avec des échantillons de voix du second chanteur[1461].

Photo

La photo fait l'objet d'innombrables solutions exploitant l'IA, à commencer par les applications photos des smartphones qui savent gérer la mise au point en analysant les scènes photographiées.

Elle sert aussi au classement de photos et à la détection de leur qualité. Les outils de gestion de bases de photos intègrent progressivement des briques d'IA les rendant un peu plus smart[1462].

Adobe ajoute régulièrement des fonctionnalités à base d'IA dans ses logiciels comme Lightroom, qui font de plus en plus appel à des ressources de calcul dans le cloud. Mais ils ne sont pas seuls sur ce créneau[1463].

Par exemple, l'application iOS **Pixelmator Photo** récupère le style d'une photo pour le répliquer automatiquement sur une autre photo. Elle utilise du deep learning entraîné sur des dizaines de milliers de photos. C'est une fonction plus intelligente que le simple copier/coller des paramètres d'ajustement d'une photo (éclairage, contrastes...). Elle peut d'ailleurs opérer à partir de la sélection d'une partie de la photo dont il faut s'inspirer[1464]. Le tout est complété d'une fonction d'upscaling, maintenant de plus en plus classique.

Il faut aussi compter avec **Meero** (2016, France, $293,4M) qui est spécialisé dans l'usage d'IA pour le traitement et la retouche automatique de photos et qui vise les marchés professionnels comme l'immobilier, le e-commerce et la restauration. Ce n'est pas anodin : cette startup qui est parmi les mieux financées dans l'IA en France. Elle gère la production de photos en s'appuyant sur des photographes externes mais toute la post-production est automatisée, ce qui permet de réduire le coût des prestations. La startup revendique avoir déjà 40 000 clients avec une répartition mondiale.

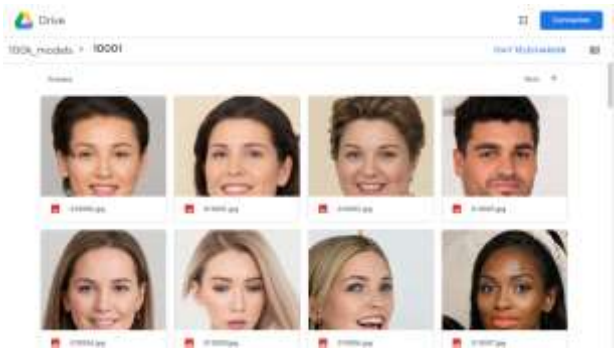
[1461] Voir Facebook's AI can convert one singer's voice into another par Kyle Wiggers, avril 2019 qui fait référence à Unsupervised Singing Voice Conversion par Eliya Nachmani et Lior Wolf, 2019 (5 pages).

[1462] EyeQ (Canada) faisait l'acquisition d'Athentech en avril 2019, dont le logiciel d'optimisation de photo Perfectly Clear qui est proposé aux utilisateurs de smartphones.

[1463] Voir Lightroom CC 2.0: What's new, and where is it headed? par Jeff Carlson, octobre 2018. Dans cette version, on retrouve la détection de visages permettant de trier les photos par personne. On peut aussi faire des recherches dans ses photos en en décrivant le contenu (ciel, mer, bâteau, avion, etc).

[1464] Voir Pixelmator Photo 1.2 update adds cursor support, ML Color Match, Split View and more par Gannon Burgett, avril 2020.

Si vous avez besoin de visages pour alimenter vos illustrations, vous pouvez maintenant faire appel aux visages générés par une IA générative, dans le site **Generated Photos** lancé par le fournisseur d'icons **Icons8**.



Ce sont des visages composites très réalistes qui ne sont pas censés correspondre à des personnes réelles. Ce qui reste d'ailleurs à démontrer. La base est disponible et téléchargeables sur Google Drive[1465].

Banuba (2016, Hong Kong, $12M) propose un SDK aux développeurs d'applications mobiles pour interpréter le contenu d'images ou les modifier. C'est une belle boîte à outils multifonctions qui vise surtout le marché des développeurs d'applications de réalité virtuelle ou augmentée. Elle sert à détecter les émotions, suivre le regard, créer des modèles 3D de visages intégrables dans des applications de réalité virtuelle, faire du mocap (motion capture, capture de mouvements, vidéo), détecter et modifier la couleur de la peau et des cheveux, améliorer les visages et enfin, enlever le fond derrière une personne.

Chez **Google**, l'application mobile Photobooth détecte automatiquement les sourires dans les visages pour capturer correctement des selfies avec plusieurs personnes[1466]. C'est maintenant assez courant dans les smartphones et appareils photos. Depuis octobre 2020, la nouvelle version de l'application Google Photos sous Android propose des suggestions diverses d'amélioration de photos exploitant du machine learning, notamment pour la gestion de l'éclairage en mode portrait, Portrait Light, qui fonctionne sur les smartphones maison Pixel 4a 5G et Pixel 5.

L'une des fonctionnalités photo majeure des smartphones est en effet le mode portrait. Celui-ci consiste non seulement à adapter l'éclairage ambiant, mais aussi à flouter son arrière-plan (effet typique obtenu en jouant sur l'ouverture d'un gros objectif photo, dont sont dépourvus les smartphones). Le floutage de l'arrière-plan est maintenant correct sur iOS comme sur Android.

Il exploite de plus en plus souvent un réseau de neurones de segmentation de l'image qui identifie les pixels qui appartiennent au visage et ceux qui relèvent de l'arrière-plan. Seul ce dernier est alors plus ou moins flouté et de manière paramétrable au niveau de la focale simulée de l'optique du smartphone avec ce que l'on appelle un flou gaussien (gaussian blur)[1467].

Le rendu peut être amélioré avec les caméras dorsales du smartphone qui peut créer une carte de profondeur de la scène. Cela peut passer par des capteurs dual-pixels qui sont exploités pour l'autofocus, par l'usage de plusieurs caméras comme dans les iPhone depuis 2016 qui créent une carte de la scène à neuf niveaux de profondeur, ou dans les Google Pixel récents, un double capteur infrarouge de profondeur pour les selfies[1468].

[1465] Voir AI Generates 100,000 Faces To Make It Available For Free par Ankush Das, septembre 2019.

[1466] Voir Google's Photobooth brings automated selfie-shooting to the Pixel 3 par Lars Rehm, avril 2019.

[1467] La méthode du floutage de l'arrière plan est bien décrite dans Synthetic depth-of-field with a single-camera mobile phone par Neal Wadah et al, Google, 2018 (13 pages).

[1468] Voir uDepth: Real-time 3D Depth Sensing on the Pixel 4 par Andrey Ignatov et al, ETH Zurich, avril 2020 (11 pages).

On attend encore une fonction plus avancée qui flouterait également l'arrière du visage comme le font les optiques à grande ouverture d'appareils photo reflex dans certaines conditions[1469]. Cela pourrait aussi exploiter la mesure de la profondeur des composantes du visage, qu'un iPhone 12 Pro pourrait réaliser grâce à son LIDAR. Pour l'instant, on n'en voit la couleur que chez Huawei.

On voit ci-dessous une photo en mode portrait réalisé à gauche avec un iPhone 12 Max Pro qui floute légèrement l'arrière des cheveux et à droite avec un Samsung Galaxy S20 qui les maintient net de bout en bout[1470].

Les Google Pixel n'ayant qu'une caméra s'appuient uniquement sur du deep learning.

Google a récemment entraîné son mode portrait avec son logiciel Portrait Light qui permet de contrôler avec précision et a posteriori l'éclairage dans le mode portrait sur ses smartphones Pixel. On peut choisir d'ajouter un éclairage directionnel. Le système a été entraîné en captant des portraits de 70 personnes dans un studio équipé de 64 caméras et 331 éclairages LED pour simuler une très grande variété d'éclairages[1471].

Dans les usages de l'IA, remarquons cet appareil photo **Alice** lancé par la startup britannique **Photogram.ai** en décembre 2020. C'est un appareil photo qui exploite votre smartphone comme écran de contrôle et outil logiciel de gestion de la prise de vue et des photos. Il est équipé d'un capteur micro 4/3 classique d'hybride de 11 mpixels à optique interchangeable et d'un processeur de traitement spécialisé pour l'IA. Il permet de bénéficier du meilleur des deux mondes : un grand capteur (mais pas au point que ce soit un plein format 24x36mm) et le support logiciel et communication du smartphone (via Wi-Fi 5 GHz). A ce stade, ce n'est qu'un prototype dont la commercialisation prévue à $1K reste hypothétique.

Vidéo et TV

Passons à la génération de vidéos à base d'IA où la créativité est aussi grande que dans celle de la musique. La plupart de ces outils font du montage automatique de vidéos et photos existantes. Les grands éditeurs du marché comme **Adobe**, **Avid** et **Elemental** (filiale d'Amazon) ajoutent fonctions après fonctions de traitement automatique des vidéos pour faire gagner du temps aux monteurs[1472].

[1469] Voir Rendering Natural Camera Bokeh Effect with Deep Learning par Andrey Ignatov et al, juin 2020 (11 pages) qui traite de progrès dans les réseaux convolutifs pour gérer ce bokeh mais qui n'arrivent visiblement pas à le générer à l'arrière d'un visage. Le modèle décrit dans ce papier est entraîné avec des photos captées avec des appareils photo reflex.

[1470] Source des photos : iPhone 12 Pro Max VS Samsung Galaxy S20 - Camera Comparison!, par Redskull sur YouTube, novembre 2020.

[1471] Voir Portrait Light: Enhancing Portrait Lighting with Machine Learning, décembre 2020.

[1472] Voir Top Video Artificial Intelligence and Machine Learning at NAB 2018, mai 2018.

Il vous est peut-être déjà arrivé de tomber sur des vidéos sur YouTube se présentant sous la forme de slideshow avec une voix off robotisée lisant un texte. Ce sont des bots de génération de spam vidéo ! Bref, le pire [1473]! Dans d'autres cas, ce sont des vidéos qui se lancent toutes seules dans des journaux en ligne et qui lisent les articles. C'est bien pour les mal voyants mais pénible pour les autres ! Heureusement, la génération automatique de vidéos a d'autres applications que voici.

Arraiy (2016, USA, $13,9M) génère des effets spéciaux automatiquement pour le cinéma, la TV et les jeux vidéo. Le principe consiste surtout à extraire par détourage des personnages du fond de l'image pour les incruster dans d'autres scènes, sans passer par l'usage d'un fond vert ou bleu.

Cela reprend le très vieux procédé du rotoscope. Le tout utilise des réseaux de neurones entraînés avec de gros volumes d'archives de vidéos. Le projet est en cours de développement.

Wibbitz (2011, Israël, $30,8M) est connu pour sa solution qui génère automatiquement des vidéos d'actualité à partir de contenus textuels et de scrapping de contenus vidéos. Elle est surtout utilisée par des médias de la presse écrite qui ont besoin de compléter leur production par des contenus vidéo dans leur version en ligne[1474]. Elle a comme concurrent direct **Wochit** (2012, Israël, $28,8M).

Magisto (2009, Israël, $22,5M) a créé un outil de montage de vidéos dans lequel on télécharge ses éléments photos et vidéos existants, qui les analyse et réalise ensuite un montage automatique à partir d'un choix préalable de style et de musique d'accompagnement. Ce sont peut-être eux qui proposent ces satanées musiques d'ukulélé aux startups IOT ! L'algorithme analyse la dimension visuelle, audio et narrative de la vidéo.

Cela cible surtout la création de vidéos à des fins de marketing (vidéo). C'est assez impressionnant et ça marche en mode web et mobile.

Graava (2014, USA, $2,3M) a créé une caméra qui fait elle-même son propre montage automatique. Et dans un registre voisin, **Soloshot** (2011, USA) a conçu une caméra robotisée pour capter automatiquement les objets en mouvement, ce qui n'a rien d'extraordinaire mais peut éventuellement s'appuyer sur de la reconnaissance d'image à base de réseau convolutif.

Triller (2015, USA, $11,9M) propose une application mobile qui peut créer une vidéo montée exploitant d'un côté de la musique de répertoire et des vidéos filmées par les utilisateurs. Il va synchroniser le tempo du montage avec celui de la musique. C'est déjà un phénomène chez les jeunes youtubeurs (vidéo). Mais c'est quelque peu déceptif et ce n'est pas qu'une question d'IA.

[1473] Voici un exemple de ce genre de vidéo : https://www.youtube.com/watch?v=Kkwq1Uht_A0.

[1474] C'est ce que fait également **Oovvuu** (2014, Australie, $3,8M) .

Remo Tech (2016, Chine) lançait en janvier 2019 une caméra automatique, l'OBSBOT Tail. Celle-ci est capable de suivre le sujet à filmer et à le cadrer automatiquement. C'est un cas d'usage direct des réseaux de neurones de type R-CNN qui détectent les contours d'objets, sans les labelliser pour autant (vidéo). Une solution voisine est proposée par **Sony** avec le REA-C1000, ou Edge Analytics Appliance, qui est un système de capture de vidéo de formateurs.

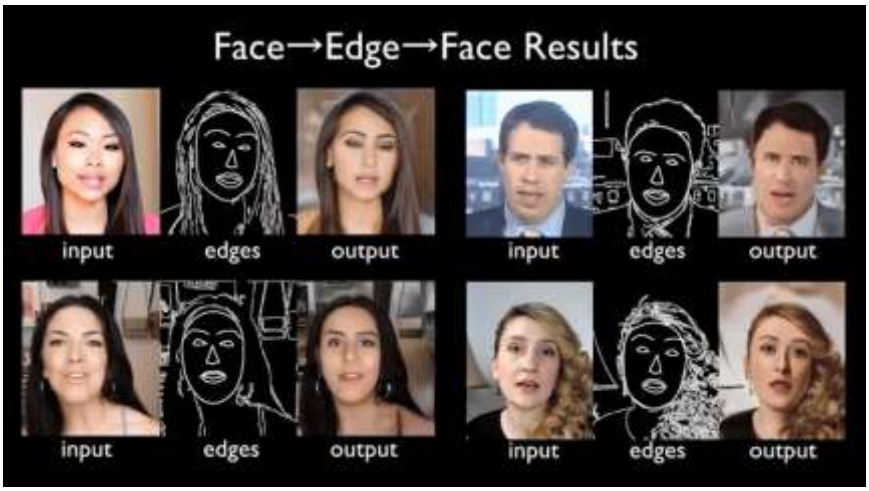
Remo Technology OBSBOT Tail

Sony REA-C1000

Il permet de faire un zoom sur le formateur et sur ce qu'il écrit au tableau le cas échéant.

Adobe Première contient maintenant aussi un outil de cadrage automatique des vidéos. Il exploite comme il se doit des réseaux de neurones probablement de type R-CNN et CNN pour identifier les objets dans les images, les labelliser et en déduire le meilleur cadrage, notamment pour respecter la fameuse règle des tiers de composition des photos qui veut notamment que les points d'attention comme les yeux d'une personne soient plutôt situés sur une ligne horizontale située au tiers haut de la photo.

Nvidia vid2vid est un réseau génératif (GAN) permettant la création de vidéos à partir de vidéos recomposées (vidéo)[1475]. Ils le font en résolution 2K. Cela fonctionne avec des vues de la route à partir de véhicules ainsi que pour l'animation de visages et de corps. Cela permet notamment de remplacer un visage dans une vidéo d'une personne en train de parler.



ScriptBook (2015, Belgique, $1,4M) ne génère pas de vidéo mais analyse les scripts de films automatiquement pour aider à les choisir avec son logiciel Script2Screen. Elle prévoit même le succès des films au box-office, ce qui est quelque peu présomptueux sans connaître le réalisateur et les acteurs. Le décalage entre un script et le produit fini d'un film est quand même généralement significatif.

Malgré tout, leur premier test ne donne aucun faux positif sur un échantillon de 65 films d'une maison de production d'Hollywood et élimine plus de scripts ayant généré des échecs que les humains impliqués dans le processus.

IBM Watson a été utilisé pour créer un système de génération automatique de bande annonce de films, testé notamment pour un film inconnu, Morgan (vidéo).

En Chine, des présentateurs d'information virtuels sont diffusés sur des chaînes d'information. Malheureusement, ils n'ont pas beaucoup d'expression[1476].

[1475] Voir Video-to-Video Synthesis, par Ting-Chun Wang & Al, de Nvidia, août 2018 (14 pages).

[1476] Voir Xinhua unveils world's first female AI news anchor, mars 2019.

Leankr (2012, France) est une startup qui génère automatiquement l'habillage de contenus passant à la TV ou de vidéo. Il les indexe et les labellise automatiquement par traitement du langage (speech-to-text, NLP) et aussi OCR (pour l'habillage TV).

Kandao Technology (2016, Chine) propose ses caméras 360° QooCam et Obsidian qui peut créer un ralenti artificiel x10 par le calcul d'images intermédiaires (vidéo). Amateurs d'émotions sportives, cela va vous enchanter !

CyberLink Perfectlink est un logiciel de gestion de web conférences qui remplace maintenant l'arrière-plan par quelque chose qui conviendra mieux que votre salon mal rangé ou votre cuisine. Ou plus simplement, il floutera tout ça. C'est aussi de l'IA, comme les fonctions photo des smartphones qui floutent l'arrière-plan de vos joli portraits et selfies. Le logiciel est aussi à même d'améliorer votre visage dynamiquement, vous permettant de faire l'économie du maquillage. Cette solution est proposée sur abonnement.

La TV et le sport sont un autre beau point d'intersection pour l'IA. Des briques d'analyse d'images sont maintenant employées pour analyser en temps réel aussi bien des matches de foot ou de tennis, entre autres sports[1477]. Cela peut servir à automatiser l'habillage TV des matches mais aussi pour aider les entraîneurs à coacher leurs équipes[1478]. Cela commence à faire son apparition dans le football, le tennis et même la gymnastique[1479] !

Tedial (Espagne) propose Smartlive, une régie vidéo dédiée au sport qui automatise l'habillage des contenus et leur diffusion dans les médias en ligne.

Minute.ly (2014, Israël, $12M) utilise le deep learning pour créer des versions très courtes de vidéos en identifiant les parties les plus intéressantes. Elles servent notamment à créer des vignettes animées pour les sites web des news ou de médias vidéo.

Le plugin **Trint** d'Adobe Premiere est capable de générer automatiquement le sous-titrage des contenus vidéo via sa fonction de speech-to-text[1480]. Après, il ne reste plus qu'à corriger les erreurs qui ne manquent pas d'être générées dans ce genre d'outil, mais cela fait tout de même gagner du temps. Premiere est aussi capable de générer une version verticale de vidéos 16/9 en cadrant automatiquement le contenu en fonction de ses zones d'intérêt.

Le projet **Imagine This!** des Universités de l'Illinois et de Washington (Seattle) génère automatiquement des dessins animés des Flintstones à partir du script[1481].

[1477] Voir Artificial Intelligence in Sports, Matias Pottala, 2018 (37 pages).

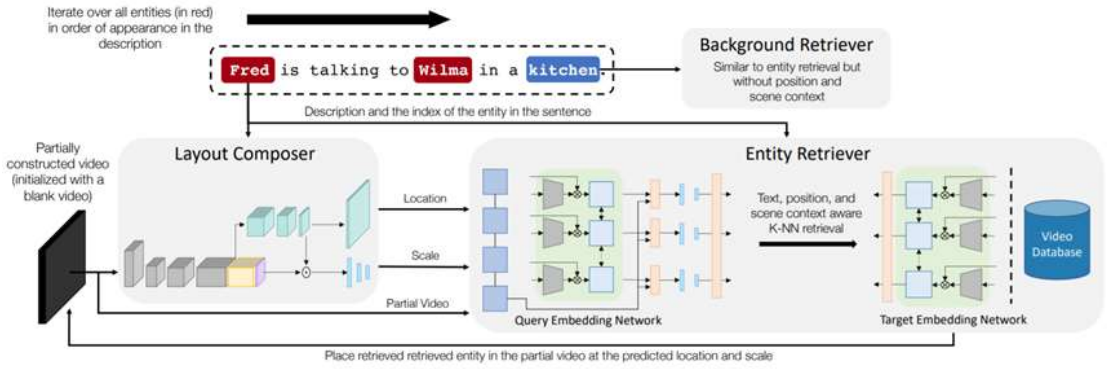
[1478] Voir Man Vs. Machine: Is Soccer Ready For Artificial Intelligence? par Robert Kidd, 2019.

[1479] Voir Fujitsu Plans to Support Professional Judges With Lidar and AI at Gymnastics Meets par Julianne Pepitone, novembre 2018.

[1480] Voir Trint's AI-powered plug-in automatically creates captions for Premiere Pro CC, mai 2018.

[1481] Voir Imagine This! Scripts to Compositions to Videos, avril 2018 (22 pages).

On est dans le registre des réseaux génératifs, c'est assez sophistiqué mais le résultat n'est pas ébouriffant pour autant car le mode d'animation généré est très simpliste dans ce type de dessin animé (vidéo).



Les réseaux génératifs opèrent aussi sur le langage et ont été expérimentés pour la création de scénarios de publicités et de film. C'est le cas du film de science-fiction **SunSpring** lancé en 2016, créé avec un système de deep learning associant des LSTM (réseau à mémoire) et RNN (réseau récurrent) entraînés avec des dizaines de scénarios de films existants de plusieurs décennies de cinéma. Une petite nouvelle sur Harry Potter a aussi été générée sur le même modèle en 2017 par **Botnik Studios** (2016, USA, $100K)[1482]. Mais l'IA n'est pas seule en jeu. Elle a suggéré des passages de textes qui ont été ensuite complétés par des écrivains. C'est donc de la créativité humaine assistée par de l'IA[1483]. Enfin, citons Morpheus de **Rct Studio** (2018, USA, $13,2M), une IA qui créé des scénarios de jeu avec un réseau génératif (vidéo)[1484].

Le machine learning est aussi mis à contribution dans la prévision d'audimat[1485] et la recommandation de contenus. Netflix en est particulièrement friand[1486].

Cinelytic (2015, USA, $2,3M) propose une IA qui analyse les bandes annonces des films pour prévoir leur résultat au box-office [1487]. Son tableau de bord d'analyse permet de faire des prédictions avant le tournage en exploitant les informations disponibles sur le casting et la thématique du film. Le tout est réalisé pays par pays en fonction des affinités connues les caractérisant. L'outil est notamment utilisé par la Warner. A ce stade, il ne semble pas qu'il soit encore capable de prévoir le succès du film en se contentant d'en lire le scénario. Cela doit être d'ailleurs plutôt difficile à faire.

[1482] Voir Harry Potter and the Portrait of What Looked Like a Large Pile of Ash, 2017.

[1483] Voir This Harry Potter AI-generated fanfiction is remarkably good par Shannon Liao, 2017.

[1484] Voir Des anciens de l'équipe du smart speaker Baidu réalisent des films avec l'AI, 2019.

[1485] Voir AI for Audience Prediction and Profiling to Power Innovative TV Content Recommendation Services par Lyndon Nixon et al, octobre 2019.

[1486] Voir How Netflix Uses AI and Machine Learning par Allen Yu, février 2019, Machine Learning - Learning how to entertain the world, Netflix et iQIYI : le Netflix chinois se sert de l'IA pour améliorer l'efficacité de son service par Valentin Cimino, décembre 2019.

[1487] Voir Artificial Intelligence (AI): Can It Help Make Hollywood Blockbusters? par Tom Taulli, janvier 2020.

Pex (2014, USA, $7M) analyse aussi le succès potentiel d'un film via sa bande annonce, mais en exploitant les métadonnées associées provenant des vues, partage et like ans les réseaux sociaux.

Art

Les arts graphiques peuvent aussi tirer parti de l'IA à plusieurs niveaux. Cela tourne beaucoup autour des techniques de reconnaissance d'image et de réseaux génératifs. Les premiers peuvent servir à reconnaître les œuvres dans les applications muséales ou pour les chercheurs. Ils servent aussi à détecter les faux et copies. Les seconds permettent d'étendre la palette créative des artistes. Cela génère des débats intéressants sur la notion de créativité attribuée aux machines. En pratique, cette créativité est toujours sous contrôle humain : celui du développeur, celui des données d'origines humaines qui alimentent l'IA et celui des humains qui font très souvent un tri sélectif de différentes options proposées par l'IA[1488].

Une IA provenant du **Rudjer Boskovic Institute** de Croatia à base de machine learning a été utilisée pour identifier les caractéristiques graphiques des peintures qui ont le plus marqué l'Humanité au travers des siècles[1489]. Elle visait à séparer l'esthétique, la valeur sentimentale et la mémorisation dans la création de la popularité. Bref, on est dans la quantification du « beau ». La solution a analysé 100 000 œuvres issues de WikiArt. Résultat ? Les peintures marquantes sont celles qui présentent le plus d'éléments saillants comme des personnages souriants ou les nus ou au contraire des figures abstraites qui n'ont pas d'équivalent dans la réalité. Mais la qualité et la vivacité des couleurs ont un impact négatif. Et l'esthétique a un impact négatif sur la mémorisation.

D'autres IA à base de GAN sont capables de générer des œuvres d'art artificielles issues du mash-up de grandes bases de données de peintures[1490]. Cela donne des résultats plus ou moins probants, ces derniers étant toutefois toujours sélectionnés par des humains. Ils génèrent ce que l'on appelle le biais du survivant : on s'ébaubit du résultat en négligeant tous les résultats décevants qui ont dû être laissés de côté. Or ce tri est une tâche éminemment humaine[1491] !

D'ailleurs, des GANs générant des tableaux à partir de photos de modèles humains auraient bien du mal à en créer qui soient convenables avec des personnes de couleur ou de type asiatique. En cause, des bases d'entraînement insuffisamment denses pour entraîner le GAN et son discriminateur [1492]! Les tableaux de la Renaissance utilisés ne présentent généralement que des visages blancs européens !

Les réseaux de neurones génératifs peuvent servir à animer des personnages de peintures. Avec une seule image ! On le faisait déjà avec une photo. La transposition à une peinture est assez immédiate. Des chercheurs du **Samsung AI Center** de Moscou ont ainsi réussi à animer Mona Lisa[1493].

[1488] Voir Can Machines And Artificial Intelligence Be Creative? par Bernard Marr, février 2020.

[1489] Voir What Can AI Tell Us About Fine Art? par Michelle Hampson, juin 2019.

[1490] Voir An AI made a $16,000 work of art, and it's actually pretty cool par Kevin Urgiles, 2018.

[1491] Voir This Picasso painting had never been seen before. Until a neural network painted it, MIT Tech Review, septembre 2019.

[1492] Voir The AI Renaissance portrait generator isn't great at painting people of color par Morgan Sung, juillet 2019.

[1493] Voir Mona Lisa frown: Machine learning brings old paintings and photos to life par Devin Coldewey, mai 2019.

Avant le deep learning, on aurait utilisé des techniques issues des effets spéciaux au cinéma en plaquant une texture du visage sur un modèle 3D en éléments finis animé par *motion capture* comme pour la princesse Leia reconstituée dans Star Wars Rogue One. Ici, le deep learning prend le relais, entraîné avec un grand nombre de visages fixes et animés. La qualité est cependant très variable (vidéo).

Dans un autre registre, le deep learning a permis à des chercheurs de reproduire des peintures existantes avec des imprimantes 3D capable d'empiler jusqu'à 10 couches de couleurs comme l'aurait fait le peintre avec son pinceau[1494].

En 2020, on découvrait le « robot artiste » **Ai-Da** capable de réaliser des œuvres d'art peintes. C'est en fait un patchwork inspiré du robot Sophia de Hanson Robotics et de système de GAN de génération de peinture et dessin[1495].

Jeux vidéo

L'IA est aussi utilisée dans les jeux vidéo et, en fait, depuis pas mal de temps. Nombre de jeux vidéo reproduisent des mondes complexes avec un grand nombre de personnages en s'appuyant sur des réseaux multi-agents qui animent des personnages respectant leur propre mode de fonctionnement et interagissant entre eux.

Des personnages de jeux à part entière peuvent être aussi animés par des IA comme l'a expérimenté **OpenAI** dans le jeu **DOTA 2** (voir la partie). D'autres jeux comme OverWatch de **Visor** conseillent le joueur pour améliorer sa pratique de jeu en l'observant, et avec des indications en temps réel.

Mais reste à rendre tout cela réaliste avec des émotions.

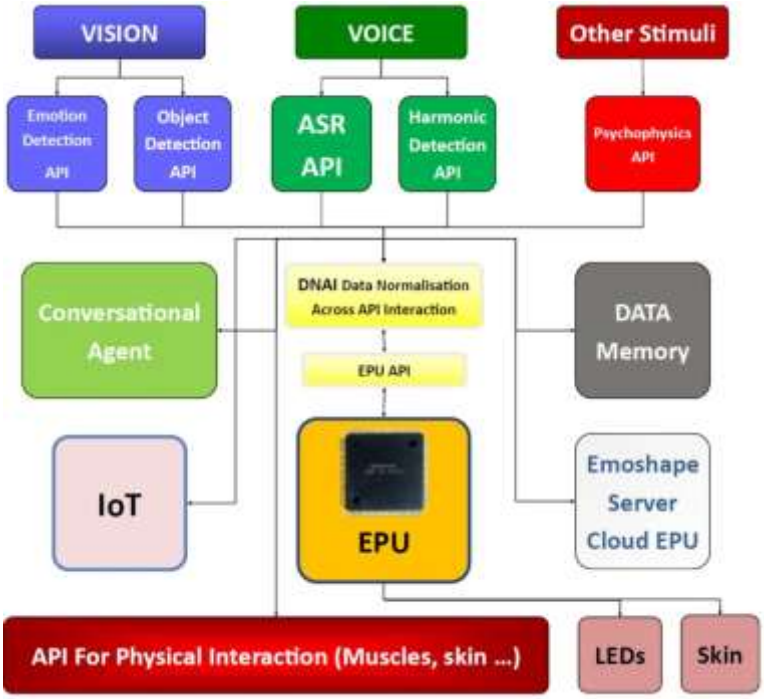
C'est ce que cherche à faire **Emoshape** (2014, USA, $370K), créée par le Français Patrick Levy-Rosenthal avec son composant électronique Emotion Processing Unit (EPU, destiné à déterminer en temps réel les émotions des utilisateurs et à permettre aux robots et autres applications comme des jeux de répondre avec un état émotionnel en phase avec celui de l'utilisateur (explication). Le chipset récupère les informations de bas niveau issues de diverses sources d'informations comme les analyses du visage réalisée par **Affectiva**, des analyses de la voix réalisées par d'autres outils et d'autres informations issues de capteurs divers (pouls...) et permet à une IA interagissant avec l'utilisateur d'adopter son propre état émotionnel, sur une palette riche de 64 trillions d'émotions différentes (vidéo), que ce soit par de la parole de synthèse comme avec WaveNet de DeepMind, de la génération d'avatars ou même la gestuelle dans le cas d'un robot humanoïde.

Le système s'enrichit de plus par l'apprentissage pour développer des états émotionnels associés aux utilisateurs qui interagissent avec lui.

[1494] Voir Researchers Use AI & 3D Printing To Reproduce Paintings par Tyler Lee, novembre 2018.

[1495] Voir This Robot Artist Just Became the First to Stage a Solo Exhibition. What Does That Say About Creativity? par Suyin Haynes juin 2019 et la vidéo d'un TEDx avec le robot Ai-Da en mai 2020 et la vidéo Robot Artist Challenges Our Definition of Art de février 2020 (5 minutes) de présentation du concept Ai-Da, créé par Aiden Meller et Lucy Seal.

Il peut par exemple être associé à un générateur de langage naturel pour lui permettre d'accentuer son intonation en fonction des interactions émotionnelles avec l'utilisateur et des textes générés par l'IA. Le chipset peut être exploité dans divers contextes : robots, enceintes vocales, jeux vidéo, etc. C'est de l'*emotion in a box*. Et ce n'est qu'un début. Ce genre de composant ou les fonctions associées seront peut-être un jour directement intégrés dans nos smartphones et laptops et leurs logiciels tiendront compte de nos émotions pour interagir avec l'utilisateur. On peut par exemple imaginer comment un moteur de recherche tiendrait compte de nos états émotionnels pour ajuster ses résultats.



Diffusion

Newswhip (2011, Irlande, $9,1M) propose un outil d'analyse de l'écho des médias et sujets dans les réseaux sociaux. Il permet d'affiner sa stratégie rédactionnelle pour que les sujets publiés collent bien aux attentes des lecteurs. Ils sont utilisés par des médias anglo-saxons comme le Huffington Post, Buzzfeed, la BBC et The Guardian.

NEWSWHIP

analytics de l'écho des médias et sujets dans les réseaux sociaux

réalise des prévisions de sujets à fort écho

à base de machine learning

références : Huffington Post, BuzzFeed, BBC, The Guardian

2011, USA, $9,1M

La startup **Echobox** (2013, UK $3,4M) propose de son côté Larry, un assistant dédié à la diffusion des contenus de médias dans les réseaux sociaux exploité par le Monde, Le Figaro, Libération, VICE et New Scientist.

Comment ça marche ? Leur IA analyse les contenus du média et les tendances dans les réseaux sociaux, puis pousse ces contenus dans la page Facebook (ou autre) du média en générant automatiquement les titres, résumés et illustrations, histoire de maximiser leur diffusion. Ça ne va pas jusqu'à choisir les illustrations pour les publier sur Instagram en fonction des photos qui sont populaires dans ce service. C'est la prochaine étape !

La recommandation de contenus musicaux ou vidéo est un gros sujet pour le machine learning et l'IA. **Netflix** améliore de manière continue ses algorithmes pour faire consommer des contenus à ses abonnés (et aussi, cacher indirectement la pauvreté de son catalogue de films)[1496].

La R&D de la **BBC** expérimentait en 2018 une IA pour optimiser la programmation d'archives vidéo sur la chaine BBC 4[1497]. **IBM Watson** est aussi appelé à la rescousse pour faire de la recommandation de vidéos sans qu'il soit évident d'évaluer les progrès en la matière[1498].

[1496].Voir Supporting content decision makers with machine learning par Melody Dye et al, décembre 2020 qui explique comment Netflix fait de la recommandation avec du transfert learning.

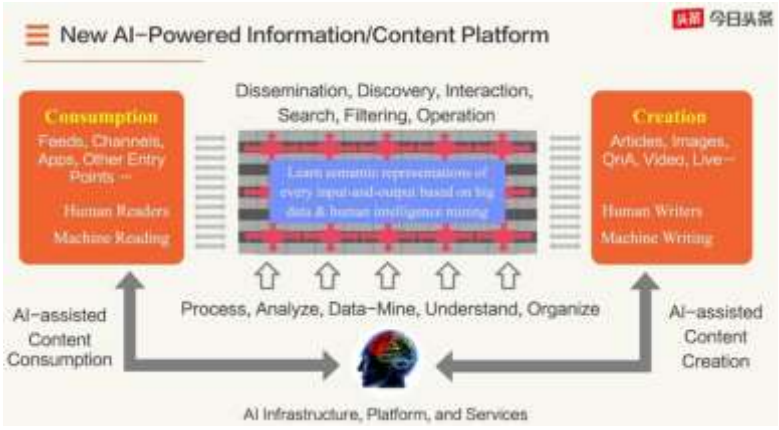
Le système analyse les vidéos que l'utilisateur a consommées au niveau audio, vidéo et métadonnées pour faire des recommandations. Le problème avec ce genre de systèmes est qu'ils ne peuvent jamais récupérer ces informations sur tout ce que l'on a pu voir depuis sa naissance. Et le risque est que si vous aimez trop les films d'actions, il ne vous proposera jamais de voir Amélie Poulain qui aurait tout de même pu vous plaire. C'est le biais du rétroviseur, une fois encore !

Enfin, cela fonctionne aussi dans la musique, comme chez **Spotify** ou **Pandora**. Et chez **Decibel Music Systems** (2010, UK) qui fournit cela son application MusicGeek pour faire de la recommandation. Et qui utilise IBM Watson. Il faut juste espérer que tout cela ne nous conduira pas à un monde ennuyeux nous faisant tourner en boucle dans nos préférences existantes sans nous faire découvrir de véritables nouveaux espaces créatifs[1499] !

Toutiao (2012, Chine, $3,1B) est un OVNI dans la sphère des startups de l'IA. Cette startup a en effet battu des records avec son financement de $3,1B et sa valorisation de $20B. Tout cela pour un agrégateur de news ! La startup est en fait une filiale du groupe ByteDance qui ne couvre que le marché chinois.

Toutiao s'appuie sur de l'IA pour sélectionner les articles et vidéos à mettre en avant pour chaque utilisateur, pour réécrire les titres des articles afin d'améliorer leur référencement et taux de click et même pour en écrire, sur le sport.

Leur audience est de 120 millions de Chinois. Ils sont aussi tentés par l'international[1500] et par la création d'un chipset d'IA sur mesure.



Monétisation

D'une manière générale, l'IA peut aider les médias à identifier les sujets porteurs en analysant les tendances dans les médias sociaux et à agencer le sommaire des médias dans leur version web et mobile.

Le californien **True Anthem** (2008, USA, $11,2M) est une plateforme intégrée de distribution de contenu destinée aux médias. Elle permet notamment l'optimisation de la distribution des contenus au travers des médias sociaux, via un ciblage de contenus assisté par IA, qui décide notamment du moment optimum pour publier les contenus.

Leur service est exploité par Reuters et CBS Interactive. D'un point de vue technique, True Anthem a l'air d'exploiter des systèmes d'analyse du langage (NLP) et des moteurs de règles.

Adomik (2012, USA, $1,3M) est une startup française qui propose un outil de prévision à base de machine learning pour optimiser la publicité programmatique. L'outil est surtout destiné aux publishers.

[1497] Voir BBC Four announces experimental AI and archive programming, août 2018.

[1498] Voir Watson Media Video Recommendations.

[1499] Voir Comment l'intelligence artificielle va rendre notre vie terriblement ennuyeuse par Céline Deluzarche, 2019.

[1500] Toutiao a fait notamment l'acquisition de l'éditeur d'applications mobiles **News Republic** (2013, France, $10,3M) en novembre 2017.

Le **Financial Times** utilise le machine learning pour déterminer la corrélation comportementale entre l'engagement dans le média et le churn[1501]. Le **Wall Street Journal** analyse son lectorat en ligne pour proposer des offres d'abonnement payant sur mesure. Une sorte de freemium automatisé par le machine learning qui exploite une soixantaine de paramètres de comportement des lecteurs[1502].

Red Points (2012, Espagne, $26,2M) détecte les copies illicites de contenus en ligne et c'est loin d'être le seul à faire cela.

Enfin, **Aive** (2017, France, 2,5M€) ambitionne d'automatiser la création et la distribution de vidéos publicitaires et promotionnelles grâce une IA « créative ». En clair, il s'agit de réaliser des montages automatiques de ces vidéos pour les adapter aux formats d'écrans et durées de vision associées, et de les sélectionner en fonction des profils des utilisateurs. La startup a été cofondée par Olivier Reynaud, ancien cofondateur de Teads, cédée à Altice en 2017 (vidéo).

Fake news

Les fake news sont devenues un véritable sujet de société, en particulier depuis que le président Trump en a fait son cheval de bataille, les dénonçant tout en générant à profusion. Les fake news n'ont cependant pas besoin d'IA pour être générées. Les Humains mal intentionnés passent leur temps à en créer et à les propager à des fins politiques ou autres. Il n'y a qu'à observer le phénomène conspirationniste QAnon qui fait feu de tout bois dans la propagation de complots les plus fous.

L'IA est devenue un instrument de cette création relativement récent, notamment avec les réseaux de neurones génératifs (GANs) apparus depuis 2014. Nous allons explorer ce thème tour à tour en nous intéressant aux différents types de fake news qui peuvent être créés avec de l'IA, les moyens de les détecter soit directement soit indirectement puis évaluer l'impact qu'elles peuvent avoir sur le métier de journaliste.

Les principaux types de fake news générés par l'IA d'aujourd'hui sont des photos, des vidéos et des textes. D'innombrables nouvelles prouesses des réseaux de neurones génératifs alimentent l'actualité depuis 2014. Ils répondent aux doux noms de Fakeapp, GAINPAINT, GPT-2, CT-GAN, FSGAN et GROVER. Ces GANs font peur. Mais pour autant, les vrais politiques, sans IA, sont aussi de très grands générateurs de fake news, avec une gradation variable selon la position dans le spectre politique !

Donald Trump en premier, qui est plus efficace qu'une batterie de 1000 serveurs de GANs avec ses plus de 30 000 mensonges et approximations proférées pendant son unique mandat de Président terminé en apothéose du mensonge sur l'élection soi-disant truquée de Joe Biden[1503].

[1501] Source : https://digitalcontentnext.org/blog/2017/06/13/artificial-intelligence-gains-momentum-news-media/.

[1502] Voir How AI powers the Wall Street Journal's dynamic paywall de Laurie Clarke, octobre 2018.

[1503] Voir President Trump has made 13,435 false or misleading claims over 993 days par Glenn Kessler, Salvador Rizzo et Meg Kelly, Washington Post, octobre 2019.

Côté photos, l'application **Fakeapp** permet de remplacer un visage par un autre dans une image. Elle permet de générer ensuite une fausse vidéo de personnalité connue[1504]. De son côté, **StyleGAN**, de Nvidia, génère des visages qui n'existent pas et il n'est pas le seul dans ce genre[1505].

Le **GAINPAINT**, un projet commun du MIT et d'IBM, permet d'éditer facilement des photos pour en altérer le contenu[1506] (vidéo). Il permet de créer des photos « staliniennes » et d'effacer facilement des personnes, des objets, ou d'en ajouter, pour peu que la base d'entraînement soit suffisamment riche. On peut même générer une vidéo d'un visage animé à partir d'une seule photo de la personne comme l'a réalisé **Samsung** en 2019 en plaquant des visages animés sur des visages d'œuvres d'art (*ci-dessous*) [1507], et aussi la voix[1508].

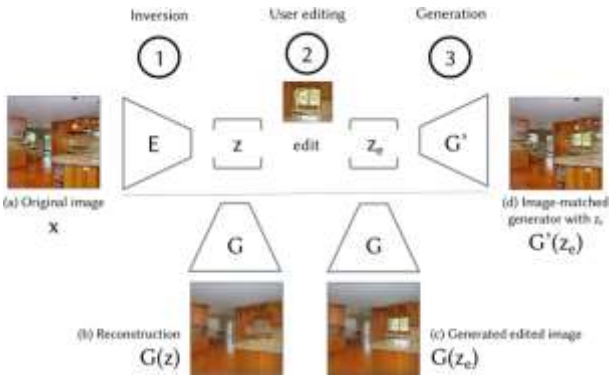


Figure 1: The results of talking head image synthesis using face landmark tracks extracted from a different video sequence of the same person (on the left), and using face landmarks of a different person (on the right). The **results** are conditioned on the **landmarks** taken from the **target** frame, while the **source** frame is an example from the training set. The talking head models on the left were trained using eight frames, while the models on the right were trained in a one-shot manner.

Quels sont les usages réels de ces outils de création de deep fake ? Un inventaire de 2019 montre qu'ils servent surtout à créer des fausses images pornographiques de femmes, et très peu d'hommes[1509]. Cela existait d'ailleurs bien avant la prolifération des GANs.

Des scénarios effrayants d'usage des GAN sur des images se font jour aussi dans l'imagerie médicale. Imaginez ainsi qu'un service d'IRM d'un hôpital soit hacké et que les images des patients soient frauduleusement altérées, provoquant une opération chirurgicale d'un patient qui soit non justifiée ! C'est l'objet de **CT-GAN**[1510].

[1504] Voir Fakeapp how Face swapping is done using ML and AI !!deepfakes tutorial, janvier 2018.

[1505] Voir This website uses AI to generate startling fake human faces par Jackson Ryan, février 2019.

[1506] Voir New AI tool lets you doctor photos, and it's showing surprising potential par Maria Dermentzi, juillet 2019 et MIT made an AI that can detect and create fake images par Christine Fisher, juillet 2019.

[1507] Voir Samsung's Deepfake AI Can Create A Video Of You Using A Single Photo par Tyler Lee, mai 2019 qui fait référence à Few-Shot Adversarial Learning of Realistic Neural Talking Head Models, mai 2019 (19 pages).

[1508] Voir New deepfake tech turns a single photo and audio file into a singing video portrait par James Vincent, juin 2019.

[1509] Voir World's First Deepfake Audit Counts Videos and Tools on the Open Web, octobre 2019.

[1510] Voir CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning, janvier 2019 (19 pages).

Une vidéo de l'imitateur Bill Hader du Late Show avec David Letterman de 2008 a été adaptée par la société **Ctrl Shift Fake**, spécialisée dans la production de deep fakes, pour ajouter les visages des acteurs imités sur celui de Bill Hader. Sa voix restait la sienne[1511].

Le réseau de neurones génératif **GROVER** de l'Université de Washington génère des fake news textuelles plausibles à 73%, ce qui n'est pas extraordinaire en soi. Mais les briques de ce réseau sont capables de détecter avec une précision de 92% les news générées par lui-même, ce qui est classique dans un GAN[1512].

OpenAI est à l'origine de **GPT**, un autre système génératif capable de générer des fake news textuelles[1513]. Il pourrait servir à créer des articles propageant de fausses nouvelles, d'imiter des auteurs et même de créer des contenus destinés à du spam ou du hameçonnage. A la base, c'est un outil de traitement du langage qui prédit le mot suivant d'un texte. Il est entraîné avec 40 Go de textes récupérés sur Internet et filtré qualitativement par des humains.

Le réseau de neurone GPT-2 comprenait 1,5 milliards de paramètres et 48 couches, entraîné avec 8 millions de ces pages web. OpenAI en avait fait la publicité en 2019 en annonçant que cette technologie ne serait pas diffusée intégralement car elle est trop dangereuse[1514]. Ils n'en ont diffusé qu'une version limitée à 350 millions de paramètres. Ce modèle n'est en fait pas très éloigné de la fonction d'auto-remplissage de texte que Google a intégré dans Gmail ainsi que dans Google Docs. Il est aussi exploitable pour faire de la traduction, mais avec un résultat du niveau d'une traduction mot à mot automatique. Ce genre d'outil pourrait servir d'assistant d'écriture dans votre traitement de textes, à créer des chatbots plus sophistiqués et à améliorer les systèmes de reconnaissance de la parole. Et puis, un an plus tard, OpenAI sortait GPT-3 avec ses 175 milliards de paramètres !

FSGAN permet de remplacer un visage par un autre dans une vidéo et en temps réel, rendant les deep fakes encore plus accessibles et dangereuses[1515] (*ci-contre*).

C'est un équivalent de vid2vid de Nvidia qui a été déjà évoqué précédemment. Ces GAN ont toujours un côté positif et un côté obscur.

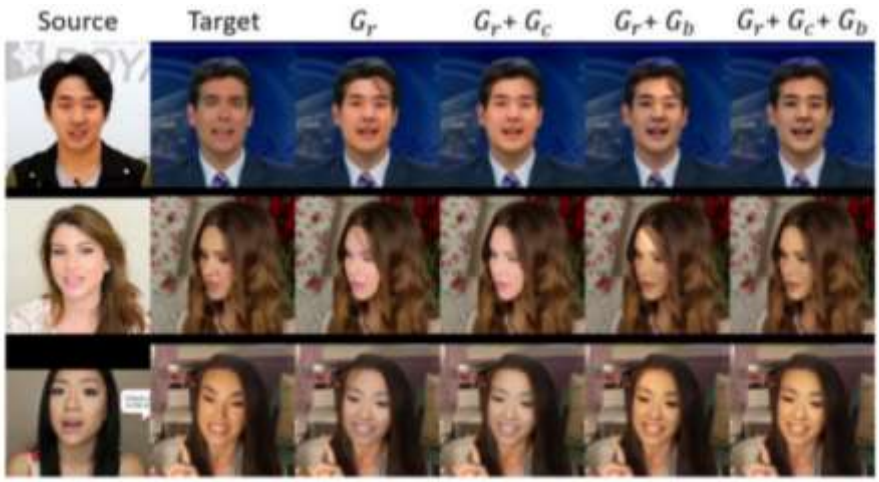


Figure 8: *Ablation study*. From columns 3 and 5, without the completion network, G_c, the transferred face does not cover the entire target face, leaving obvious artifacts. Columns 3 and 4 show that without the blending network, G_b, the skin color and lighting conditions of the transferred face are inconsistent with its new context.

Ainsi, cet outil d'édition de texte dans une vidéo qui permet de faire dire ce que l'on veut à quelqu'un est-il évidemment dangereux si mal utilisé, mais très utile pour créer un doublage avec synchronisation des lèvres[1516].

[1511] Voir Deepfake of Bill Hader doing an impression of Tom Cruise will make your brain hurt par Brian Koerber, août 2019.

[1512] Voir New AI Generates Horrifyingly Plausible Fake News par Dan Robitzskimay, mai 2019 qui fait référence à Defending Against Neural Fake News, mai 2019 (20 pages).

[1513] Voir Better Language Models and Their Implications, février 2019, qui fait référence à Language Models are Unsupervised Multitask Learners (24 pages).

[1514] Voir The AI Text Generator That's Too Dangerous to Make Public par Tom Simonite, février 2019 et Musk-backed AI group: Our text generator is so good it's scary par Marrian Zhou, février 2019.

[1515] Voir FSGAN: Subject Agnostic Face Swapping and Reenactment par Yuval Nirkin, Yosi Keller et Tal Hassner, août 2019.

[1516] Voir Text-based Editing of Talking-head Video, juin 2019 (8 mn) qui fait reference à Text-based Editing of Talking-head Video.

Encore plus bluffant, **Speech2Face** du MIT est un réseau de neurones capable de reconstituer une photo d'une personne en analysant sa voix. Cela passe évidemment par l'exploitation d'une base d'entraînement avec des millions de vidéos glanées sur Internet[1517]. Ce même réseau de neurones est adaptable pour plaquer une voix sur un visage et l'animer.

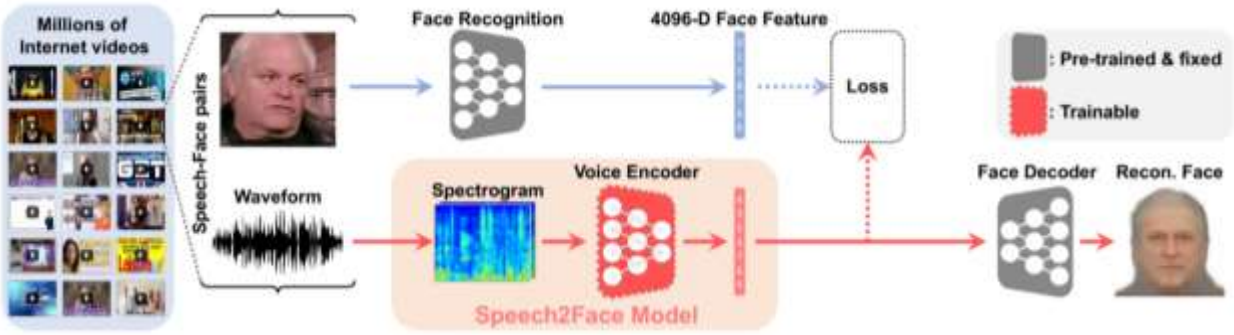


Figure 2. **Speech2Face model and training pipeline.** The input to our network is a complex spectrogram computed from the short audio segment of a person speaking. The output is a 4096-D face feature that is then decoded into a canonical image of the face using a pre-trained face decoder network [10]. The module we train is marked by the orange-tinted box. We train the network to regress to the true face feature computed by feeding an image of the person (representative frame from the video) into a face recognition network [40] and extracting the feature from its penultimate layer. Our model is trained on millions of speech–face embedding pairs from the AVSpeech dataset [14].

Passons aux usages utiles, **Overdub** de la startup **Descript** (2017, USA, $50M) permet d'éditer un podcast en modifiant le script du transcript de l'audio. Il ne demande que très peu de données d'entraînement de l'intervenant (vidéo). Le processus est simple : on enregistre son podcast, l'outil génère un transcript. Celui-ci est éditable pour corriger d'éventuelles erreurs (on s'est trompé sur un chiffre, une date, un nom). L'outil régénère alors l'audio et la voix des intervenants à partir du texte édité. C'est pratique, mais cela permettrait aussi de générer de fausses interviews ! Sachant que Descript a fait en septembre 2019 l'acquisition de **Lyrebird** qui génère de la voix artificielle. Ainsi, la boucle est bouclée.

Autre forme de fake news, les présentateurs chinois synthétiques de journaux télévisés. Comme leurs modèles humains manquent souvent d'expression, la différence entre ces derniers et les versions synthétiques ne sont pas énormes et cela peut passer. Dans les deux cas, les présentateurs ne véhiculent aucune émotion[1518].

Que faire face aux fake news qui menacent le fondement même de nos démocraties[1519] ? **Eric Scherer** résume bien les approches[1520].

[1517] Voir Speech2Face: Learning the Face Behind a Voice, 2019 (11 pages).

[1518] Voir Experts cast doubt on whether China's news anchor is really A.I. par Lucy Handley, 2018, This is fake news! China's 'AI news anchor' isn't intelligent at all, novembre 2018. Et les vidéos : Xinhua's first English AI anchor makes debut et China's female AI anchor a newsreader officially starts work, mars 2019.

[1519] Voir Les deepfake, une menace à la sécurité nationale?, juin 2019, Fake news, IA & Information : la cote d'alerte est-elle en train d'être franchie ?, juin 2019.

[1520] Voir #Fake, #DeepFake - On n'a encore rien vu ! Ne croyez plus rien de numérique ! par Eric Scherer, juin 2019.

Il faut s'entraîner à ne plus croire ce qui est créé de manière numérique, faire attention aux fausses nouvelles propagées par des groupes politiques extrémistes, se considérer mieux armé dans les pays dotés de médias publics et privés indépendants, tout questionner, éduquer les enfants, demander des comptes aux médias, renforcer les outils des commissions électorales des démocraties et, finalement, savoir douter.

Certains politiques souhaitent même légiférer pour punir la création de deepfakes, comme des élus démocrates ont essayé de le faire en juin 2019 à la Chambre des Représentants US en 2019[1521]. La proposition de **Deepfake Accountability Act** était encore étudiée en commission en octobre 2019. Ce qui veut dire qu'elle a été enterrée. De son côté, la Chine n'a pas hésité à légiférer et à rendre la diffusion de fake news vidéos illégale et pénalisée[1522].

Se pose aussi la question de la propagation des fausses nouvelles sur les réseaux sociaux. À ce jour, **Facebook** a refusé de supprimer celle de vidéos truquées sur Nancy Pelosi ou de publicités mensongères des Républicains sur Joe Biden. Cela s'est retourné contre son fondateur Mark Zuckerberg lorsqu'une fausse vidéo de lui a été diffusée sur son propre réseau[1523].

Quels sont les **outils de détection automatique de fake news**, qu'elles soient d'origine humaine ou créées par de l'IA ? On en voit apparaître de deux types : ceux qui les détectent en examinant leur contenu et ceux qui le font en examinant leur origine de création et leurs processus de propagation, en gros, via leurs métadonnées.

Il existe ainsi des outils comme **FaceForensics** et **MesoNet** qui détectent les visages altérés à partir de grandes bases de données de visages d'origine[1524].

L'une des approches, déjà vue pour GROVER, consiste à générer des deep fakes et à utiliser l'IA pour les détecter. Elle est facile à entraîner puisqu'elle génère ses propres jeux de données d'entraînement. C'est une méthode testée par **Facebook** et **Google** dans l'identification de fausses vidéos. Facebook a même lancé un concours de détection de deep fakes fin en 2019 en fournissant un jeu d'entraînement de 100 000 clips vidéos, certains normaux et d'autres altérés par des GANs. Les résultats obtenus en 2020 montraient que les deep fake « venant de nulle part », sans données d'entraînement associées, n'étaient reconnus qu'à 65%, ce qui est assez moyen. Il reste donc fort à faire[1525].

Adobe propose de son côté un système de détection de photos truquées ou simplement modifiées avec Photoshop[1526].

Du côté des textes, des outils d'analyse linguistique voient le jour pour détecter les fausses nouvelles comme **GTRL** (Giant Language Model Test Room), créé par le MIT et Harvard[1527].

Des startups se sont lancées sur le marché de la détection de fake news. C'est le cas de **Deeptrace** (2018, Pays-Bas) qui détecte les fausses vidéos dans le cadre de campagnes électorales[1528].

[1521] Voir DeepFakes Accountability Act would impose unenforceable rules — but it's a start par Devin Coldewey, juin 2019.

[1522] Voir China Bans Deepfakes In New Content Crackdown par Emma Woollacott, novembre 2019.

[1523] Voir A Fake Zuckerberg Video Challenges Facebook's Rules par Cade Metz, juin 2019.

[1524] Voir FaceForensics: A Large-scale Video Dataset for Forgery detection in Human Faces, 2018 (21 pages) et MesoNet a Compact Facial Video Forgery Detection Network, 2018 (7 pages).

[1525] Voir Facebook just released a database of 100,000 deepfakes to teach AI how to spot them par Will Douglas Heaven, juin 2020, Facebook AI Launches Its Deepfake Detection Challenge par Eliza Strickland, décembre 2019 et Facebook contest reveals deepfake detection is still an "unsolved problem" par James Vincent, juin 2020.

[1526] Voir Adobe développe une IA capable de détecter les images photoshopées par Arthur Vera, juin 2019. C'est devenu le projet About Face d'Adobe (vidéo). Le prototype indique un % de véracité d'une image, propose une heatmap des modifications réalisées dans la photo et même de retrouver l'original. Voir Adobe previews an AI feature that can tell when an image has been manipulated par Dami Lee, 2019.

[1527] Voir One potential route to flagging fake news at scale: Linguistic analysis par Fatemeh Torabi Asr, août 2019 et AI now can spot fake news generated by AI par Shelby Brown, août 2019.

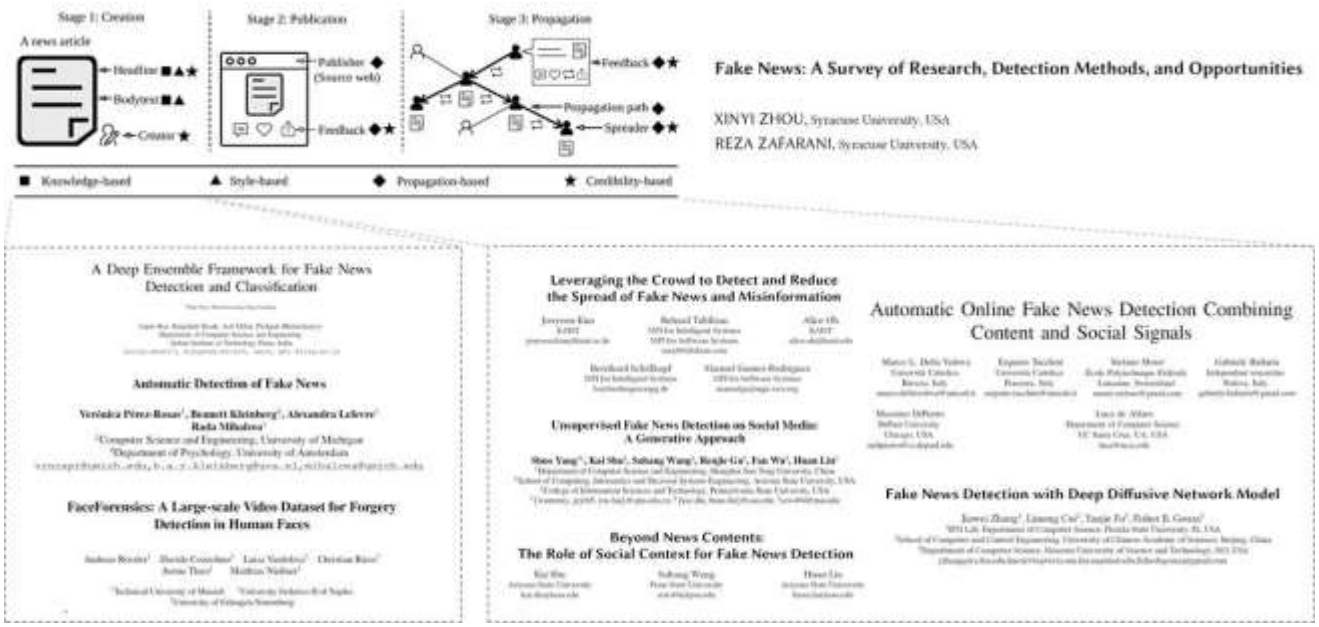
D'autres startups sont focalisées sur la détection de fake news affectant des marques, comme **New Knowledge** (USA, $13M) et **AI Foundation** (USA, $27,5M) et **AdVerifai** (2017, USA).

L'équipe du Medialab de Denis Teyssou à l'**AFP** a mené plusieurs projets qui aident les journalistes de l'agence à détecter les fake news en recherchant notamment l'origine exacte de photos et vidéos qui peuvent ne pas refléter les événements qu'ils sont censés décrire[1529].

Mais cela ne fonctionne pas forcément et le discernement humain restera nécessaire encore long-temps[1530] en particulier pour les fausses nouvelles exprimées sous une forme négative.

Au même titre qu'il est toujours difficile de prouver que quelque chose n'existe pas[1531]. On peut aussi entraîner les Internautes à reconnaître de faux visages en leur montrant des vrais et des faux visages[1532].

L'autre grande méthode de détection des fake news consiste à analyser leur origine et les sites et personnes qui les propagent[1533]. Au même titre que si vous découvrez une méthode pour guérir du cancer dans un publirédactionnel propagé par Outbrain ou Taboola, vous devez vous méfier !



Ainsi, créé en France par des élèves de l'EPITA, **Neutral News** détecte les « fake news » via leurs métadonnées d'origine et de diffusion[1534].

[1528] Voir Can AI Detect Deepfakes To Help Ensure Integrity Of U.S. 2020 Elections? par Jeremy Hsu, février 2019.

[1529] Voir AFP Medialab: information verification and user experience at the core of newsroom innovation et l'intervention de Denis Teyssou dans une table ronde au Web2day de juin 2018 à partir de la seizième minute.

[1530] Voir Facebook is making its own deepfakes and offering prizes for detecting them par Devin Coldewey, septembre 2019 et Even the AI Behind Deepfakes Can't Save Us From Being Duped, février 2019.

[1531] Voir Le Machine Learning ne peut détecter les Fake News selon une étude par Bastien L, octobre 2019.

[1532] Voir Can you tell the difference between a real face and an AI-generated fake? par James Vincent, mars 2019.

[1533] Voir par exemple Fake News: A Survey of Research, Detection Methods, and Opportunities par Xiniy Zhou & Al, décembre 2018 (d'où vient le schéma du haut de l'illustration de cette page), Unsupervised Fake News Detection on Social Media: A Generative Approach, Shuo Yang & Al, 2019 (8 pages) et Automatic Online Fake News Detection Combining Content and Social Signals, mai 2018 (8 pages).

[1534] Voir Neutral News, l'outil qui tracke les « fake news » grâce à l'intelligence artificielle, décembre 2018.

A contrario, le projet **DeepNews** lancé par Frédéric Filloux à partir de ses travaux réalisés à Stanford visent à référencer les articles de qualité pour aider à les monétiser, et à partir de leur contenu et de leurs métadonnées de propagation. C'est une approche inverse de la détection des fake news. L'outil est surtout destiné à des plateformes d'agrégation[1535].

En 2016, une analyse des tweets de Donald Trump pendant la campagne électorale permettait d'identifier leur origine en fonction des mots utilisés et de l'heure d'émission entre lui et son équipe de campagne. À l'époque, les siens provenaient d'Android et ceux de son équipe d'un iPhone[1536]. Pendant sa présidence, ses pratiques avaient changé. Il s'est mis à tweeter de manière de plus en plus compulsive, et bien plus fréquemment que son équipe. Google avait analysé automatiquement l'impact des tweets de Donald Trump et leur couverture visuelle[1537]. Il faut bien cela ! Trump est à lui tout seul un stakhanoviste de la génération et de la propagation de fake news, souvent issues de milieux conspirationnistes comme QAnon.

Cela a amené une équipe de chercheurs californiens à créer une solution d'IA à base d'analyse de graphes à même de séparer les thèses conspirationnistes des informations plus solides. Ils ont découvert que les informations solides s'appuyaient sur des réseaux d'informations plus denses tandis que les thèses conspirationnistes étaient reliées entre elles et de manière moins dense[1538]. Ceci étant dit, un peu de jugeotte intellectuelle permet de faire le tri rapidement. Malheureusement, elle n'est pas si commune que cela.

Ce phénomène des fake news et l'IA en général déstabilisent le métier du journalisme qui doit s'adapter à la nouvelle donne. Ceux-ci doivent maîtriser les outils de recherche et de validation de l'information. Ils doivent être encore plus méfiants des sources, même celles qui sont officielles comme l'a montré la fausse découverte en Ecosse de Xavier Dupont-Ligonnes en octobre 2019.

Les agences de presse s'équipent d'outils à base d'IA pour détecter les informations pertinentes comme Lynx Insight chez **Reuters**[1539]. Enfin, la **BBC** utilise l'outil Juicer pour extraire des informations pertinentes de nombreux flux RSS[1540].

Services en ligne

Les solutions sur Internet font évidemment abondamment appel aux techniques de l'IA et nous en avons déjà cité un bon nombre, ne serait-ce que pour le commerce en ligne et la publicité.

L'une des plus connues est la mécanique de recommandation de contenus vidéo de **Netflix** qui s'est améliorée sur une période de plus de 10 ans, et exploite différentes techniques de machine learning.

[1535] Voir A progress report on Deepnews.ai, mars 2018 et Deepnews.ai, Progress Report #2: It works, octobre 2018.

[1536] Voir Text analysis of Trump's tweets confirms he writes only the (angrier) Android half, par David Robinson, 2016.

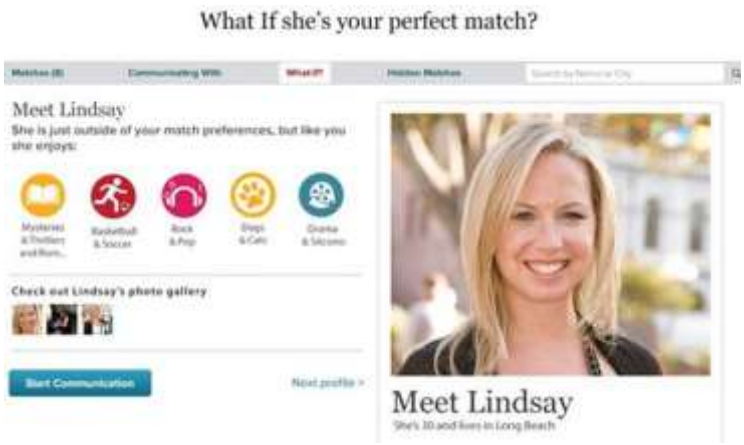
[1537] Voir Google's Vision AI Found Two Hours Of Trump's Tweets In A Week Of Television News par Kalev Leetaru, juin 2019.

[1538] Voir An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web par Timothy R. Tangherlini et al, juin 2020 (39 pages).

[1539] Voir Reuters is taking a big gamble on AI-supported journalism par Nicole Kobie, mars 2018.

[1540] Voir The Juicer.

La recommandation est mise à toutes les sauces sur Internet. C'est le cas pour les sites de rencontre comme **eHarmony** (1999, USA, $121M, anciennement Compatible) qui utilise le machine learning pour améliorer le matching proposé. Ce n'est pas très romantique ni sérendipitesque mais cela fonctionne peut-être. Ont-ils accès à un historique qui va juste qu'aux prononciations de divorces ? Il faudrait en disposer pour faire du prédictif bien documenté[1541] !



Bodyguard.ai (2017, France) protège les utilisateurs de réseaux sociaux contre les contenus et commentaires haineux et le harcèlement. Cela repose sur des briques de deep learning de traitement du langage.

Tourisme

Le tourisme est un autre terrain de jeu propice aux innovations à base d'IA : les données sont abondantes, notamment via les intermédiaires de la réservation en ligne et chez les compagnies aériennes et leurs services de réservation mutualisés tels que **Sabre** et **Amadeus**[1542]. C'est aussi un marché très grand public qui peut exploiter les outils de la mobilité et ceux des objets connectés.

Les systèmes de réservations de voyages exploitent toutes les techniques imaginables de « yield management » pour les remplir aux prix les plus élevés. Certains systèmes exploitent de la logique floue, d'autres du machine learning. **PriceMoov** (2016, France, 3M€) optimise ainsi les tarifs de différents services comme la location de véhicules (Rentacar), de chambres d'hôtel, de location d'équipements ou de billetteries en ligne. Leur "Dynamic Pricing" propose des prix dynamiques permettant d'optimiser le CA en s'appuyant sur l'état de la demande, les prix pratiqués par la concurrence et aussi la météo. Le tout exploite la solution logicielle de **Dataiku**. Il n'est pas sûr que cela améliore la satisfaction des clients finaux pout autant !

Chatbots

Les chatbots de préparation de voyages sont très nombreux, surtout aux USA. On en trouve qui sont attachés à des niches variés ou à des offres spécifiques comme pour les trains **Amtrak**, pour le processus de checking en ligne de **KLM**[1543], **Air France KLM** qui a un chatbot pour Facebook Messenger, avec le chatbot de **Voyages SNCF** sur Facebook Messenger (vidéo), dans le groupe **Accor Hotels,** toujours sur Facebook Messenger[1544] pour gérer les réservations dans un millier d'hôtel, soit le quart des hôtels du groupe, le tout étant couplé à un système de « Smart Pricing » (yield management), ou **Ask Mona** (2017, France), un chatbot de sélection de visites culturelles en France.

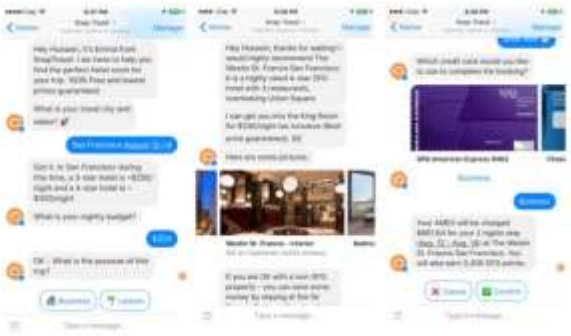
[1541] Voir eHarmony: How machine learning is leading to better and longer-lasting love matches, mars 2018.

[1542] Voir ce qu'ils en pensent dans Defining the future of travel through intelligence d'Amadeus, 2017 (22 pages).

[1543] Voir cette liste de quelques dizaines de chatbots : https://www.30secondstofly.com/ai-software/ultimate-travel-bot-list.

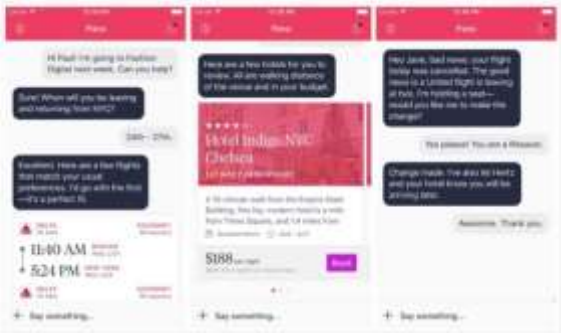
[1544] Que j'ai testé et qui n'apporte pas grand-chose, et en plus est très lent.

SnapTravel (2016, USA, $9,2M) permet de choisir son hôtel en fonction de ses contraintes budgétaires et via divers supports de communication (SMS, Facebook Messenger et même Slack). Il associe comme certains chatbots de l'IA et de l'intervention humaine et scanne les offres d'Expedia, de Priceline et de dizaines de sites.



Bref, c'est un moteur de recherche à commande textuelle.

Pana (2015, USA, $1,5M) est un équivalent destiné aux voyages professionnels. **TravelAppeal** (2014, Italie, $7,1M) est un chatbot pour Facebook Messenger pour gérer son séjour en temps réel et obtenir des réponses aux questions courantes pendant le voyage.



Le chatbot s'alimente en aspirant les contenus de 500 sources Internet dont les réseaux sociaux et sites liés au voyage et au tourisme.

Le même principe est en place depuis début 2017 dans l'hôtel **Cosmopolitan** de Las Vegas qui utilise son chatbot Rose (*ci-dessous*) fonctionnant notamment via SMS qui permet de passer toutes sortes de commandes. Ce concierge autonome aurait permis d'augmenter de 39% les commandes d'extras par les clients et 90% d'entre eux y feraient appel.

L'hôtel qui est relativement récent comprend 3000 chambres. Le chatbot a été développé par l'agence américaine **R/GA**. Vous pouvez tester le chatbot en appelant le + 1 702 930 8188 !

Les applications mobiles généralistes chatbot de préparation de voyage ne manquent pas aux USA. Nous avons par exemple **Lola Travel Company** (2015, USA, $44,6M), créé par Paul English, le fondateur du moteur de recherche de voyages **Kayak** revendu à Priceline pour $2B**, Mezi** (2015, USA, $11,8M) et **Skyscanner** (2003, USA, $197M, acquis par le chinois **Ctrip** en novembre 2016). Toutes ces applications se ressemblent et accèdent généralement aux mêmes sources de données[1545].

On commence aussi à voir des hôtels installer des systèmes à commande vocale dans les chambres. C'est le cas au **Encore** et au **Wynn** à Las Vegas dont les chambres sont toutes équipées de systèmes **Amazon** Echo pour piloter, essentiellement… la lumière.

La commande vocale peut aussi servir à effectuer sa réservation, cette fois-ci à partir de chez soi, avec ces enceintes connectées à commande vocale dont le taux de pénétration dans les foyers dépasse maintenant celui des weareables, aux USA.

Mais les statistiques d'usage semblent montrer que, pour l'instant, les consommateurs sont encore hésitants à passer des commandes via la voix et Alexa/Echo, préférant le faire dans des sites web ou applications mobiles, qui permettent de plus facilement vérifier que l'on ne s'est pas trompé. C'est peut-être aussi lié à un phénomène d'accoutumance lent à se produire comme cela a été le cas sur le commerce en ligne aux débuts de l'Internet.

[1545] Voir Artificial Intelligence (AI) in the Travel Industry, avril 2017.

Enfin, **Smartbnb** (2016, France) propose un agent conversationnel destiné aux voyageurs faisant appel à Airbnb et HomeAway. La solution identifie les questions les plus courantes des utilisateurs et apporter les réponses adaptées.

Parcours touristiques

La création de parcours touristiques personnalisés devrait être un bon champ d'application de l'IA. On indiquerait sa ville, ses préférences en termes de types de visite, le nombre de jours, les moyens de transports préférés et le budget et le système produirait automatiquement des propositions d'agendas avec horaires, transports et un forfait pour tout payer.

Mais aujourd'hui, c'est encore plutôt du domaine de la science fiction et de la singularité. Pourquoi donc ? Parce qu'il est très difficile d'obtenir toutes les données structurées nécessaires, que les marchés sont fragmentés, que les billetteries en ligne ne sont pas standardisées pour les visites et qu'il faut tout refaire dans chaque ville. Mais cela arrivera bien un jour. Vous objecterez avec tel ou tel service qui existe déjà, mais vous rendrez compte, de près, qu'il manque toujours quelque chose comme la création d'un parcours qui intègre les temps de transport.

Mais l'application mobile **Google Trip** qui fonctionne même en mode déconnecté commence à s'en approche tout de même (vidéo). Mais la vidéo de présentation est trompeuse ! Les parcours de visite sont préformatés et ne sont pas personnalisables.

Wayblazer (2014, USA, $8,7M) propose des recommandations contextualisées et personnalisées qui s'appuient sur les outils logiciels d'IBM Watson (vidéo). Il personnalise l'accompagnement photographique des propositions en fonction des recherches textuelles multicritères de l'utilisateur.

C'est une sorte de concierge numérique commercialisé aux professionnels du tourisme. Une solution équivalente est proposée par **GoMoment** (2010, USA).

Gogobot (2010, USA, $39M) et son application Trip.com utilise un modèle prédictif qui exploite la segmentation socio-démographique du voyageur, le moment et la météo pour proposer des visites. Mais l'intégration n'est pas extraordinaire au premier abord lorsque l'on teste le site qui sépare hôtel et avion alors que l'offre devrait être intégrée, comme dans **Opodo**. La startup a été également acquise par le Chinois **CTrip**.

Et des guides de visites en réalité augmentée, qui seraient des équivalents de **Pokemon Go** servant à quelque chose ? Cela arrive au compte goutte, mais avec des coûts de production par attraction qui sont encore trop élevés pour être généralisés. Reste aussi à inventer une IA qui rendrait les serveurs des cafés parisiens plus sympas et orientés clients !

Expérience touristique

L'expérience touristique peut s'améliorer en tirant parti de l'IA à différents étages.

J'ai pour l'instant découvert cet outil de prévision proposé aux hôteliers par la startup française **Victor&Charles.**

Elle s'appuie sur IBM Watson et exploite toutes vos données publiques des clients disponibles dans les réseaux sociaux pour en analyser l'influence, les affinités et l'humeur. Il propose alors des recommandations à l'hôtel qui vous accueille pour lui permettre de personnaliser votre arrivée, et notamment de trouver la personne la plus appropriée pour s'en charger.

Cela s'appliquera plutôt à des hôtels quadri-étoilés ou plus ! La solution utilise IBM Watson Conversation, Natural Language Understanding, Personality Insights et Tone Analyzer. Le matching de personnalité est une fonctionnalité que la startup souhaite commercialiser au-delà du marché de l'hôtellerie. Son développement n'a pas duré plus de deux mois.

Et puis nous avons cette expérience client en environnement fermé présentée par **Carnival** lors d'un keynote au CES 2017 (vidéo). Elle consiste à proposer un badge RFID aux passagers des paquebots qui permet d'accéder à tous les services du navire, ces services étant personnalisés en fonction des préférences et de l'activité des passagers. Le tout avec force machine learning exploitant l'historique de consommation des passagers.

Depuis quelques années, vous pouvez installer sur votre smartphones diverses applications, notamment de **Google**, qui traduisent automatiquement la signalétique tout comme les menus de restaurants.

Il y a aussi fort à faire du côté des agents vocaux pour accompagner les touristes. C'est ce que fait **Aiello** (Taïwan). Et puis aussi côté chatbots avec **Mindsay** (2016, Paris). La commande vocale dans certains restaurants commence aussi à faire son apparition.

Robots

Si vous aimez les robots, le tourisme pourra vous donner l'occasion d'en croiser, mais plutôt rarement, et surtout en Asie, le continent qui n'a pas peur des robots.

Tout d'abord en vous équipant vous-mêmes d'une valise robot comme l'étonnante **Cowarobot** (vidéo). D'origine chinoise, la startup avait réussi sa levée de fonds sur IndieGogo avec $581K de récoltés en septembre 2016. A ce jour, les valises ne sont toujours pas livrées aux early-adopters !

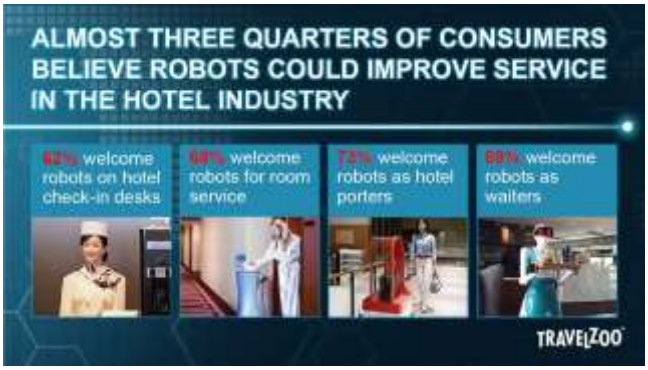
Vous pouvez aussi croiser des robots mobiles d'information de **Qihan** dans les aéroports comme à Shanghai. **Starwood** a mis en service un robot majordome à roulettes et tablette dans un hôtel à Palo Alto en 2014. Il permet de livrer dans les chambres les petits ustensiles demandés par les clients, comme des serviettes, des brosses à dent ou des accessoires électriques. Il provenait de **Savioke** (2013, USA, $34M) qui est de la même région. Ce robot est aussi déployé dans des hôpitaux. Le cours de l'Histoire aurait été changé si un tel robot avait été installé dans le Sofitel de New York en 2011 !

Enfin, si vous allez au Japon, vous pouvez faire un séjour dans l'hôtel pilote **Henn-na** près de Nagasaki avec ses 75 chambres et dont l'accueil et le service étaient réalisés par des robots depuis 2015 (vidéo).

En 2019, l'hôtel congédiait la moitié de ses robots qui ne fonctionnaient pas bien et coûtaient trop cher en maintenance. Mais l'autre moitié est toujours là, notamment les robots de l'accueil[1546].

Vous avez le choix entre une hôtesse robot ou un vélociraptor robot qui ne font que servir d'interface visuelle pour l'automate qui vous permet de faire votre checking et qui existe déjà dans diverses chaines d'hôtel en France. Ce même hôtel robotise le transport de vos bagages dans votre chambre. Dans des cas relevant moins de la science fiction, certains hôtels commencent à installer des bornes de check-in avec reconnaissance faciale. Ce n'est qu'un début ! Après, si vous avez besoin de rencontrer de vraies personnes pour alimenter votre organisme en sérotonine et en dopamine, c'est une autre affaire !

[1546] Voir Au Japon, l'hôtel géré par des robots fait machine arrière par Rafaële Brillaud, janvier 2019.



Mode et luxe

Les marchés de la mode et du luxe sont hétérogènes mais partagent un certain nombre d'aspects pour ce qui est des cas d'usages de l'IA. Ils sont, pour une part, voisins de ceux du retail, du fait des réseaux de point de vente des marques concernées. Les marques grand public peuvent aussi adopter des chatbots pour fluidifier leur relation clients.

Les sites web peuvent être enrichis de fonctions de recherche de produit par similarités visuelles. L'IA peut aussi intervenir dans le processus d'étude de tendances, dans la création ainsi, en aval, que dans la détection de contrefaçons.

La présence d'un chercheur en IA dans une publicité de parfum d'Yves Saint Laurent est anecdotique mais symbolique de la valorisation de ce genre de profil professionnel. Ces secteurs d'activité sont très traditionnels. Ils s'appuient sur des jeux de données de volume variable selon le niveau d'élitisme des produits commercialisés. Le fashion grand public est un marché de volume fragmenté. Le luxe est tout aussi fragmenté mais génère moins de volume.

Création

L'IA peut être impliquée dans la conception de produits ainsi que dans les unités de production, surtout celles qui sont industrielles.

Cela peut aller jusqu'à la création de produits avec l'aide d'IA. En complément des outils de conception 3D (CAO), le principe consiste à utiliser des réseaux de neurones génératifs (GANs) pour créer des modèles à partir de modèles existants. Le débat sur la créativité de l'IA va bon train dans un pareil cas. En pratique, les créatifs humains jouent avec la palette qui est proposée par les logiciels et retiennent ce qui leur semble le plus intéressant. Ces IA restent des outils puissants à disposition des créatifs plus que des outils remplaçant les créatifs. Comme un traitement texte pour un auteur ou un logiciel de CAO pour un ingénieur.

L'un des exemples les plus connus est l'artiste et chercheur en IA Robbie Barrat de **Balenciaga** qui s'est ainsi appuyé sur une IA générative pour créer une collection complète s'appuyant sur des modèles existants de la marque[1547].

L'IA créait des variantes de styles et couleurs à partir de l'existant. Le résultat est assez brouillon[1548].

[1547] Voir Can an AI do Balenciaga better than Balenciaga?, septembre 2018.

[1548] Voir https://threadreaderapp.com/thread/1021476460345708544.html. Notons à contrario que la génération de peintures artificielles fonctionne assez mal si l'on cherche à obtenir un rendu réaliste. C'est le cas avec AI is used to automatically create nude portraits.. and the results are horrifying, mars 2018, qui évoque le cas de GANs générant des peintures de nus.

Est-ce qu'une IA peut détecter les tendances en amont de leur apparition via une détection de signaux faibles[1549] ?

C'est une promesse facile à faire mais plutôt difficile à réaliser. Les bases d'entraînement ne sont pas évidentes à constituer. Les signaux faibles sont de nature diverse et pas évidents à intégrer. Qui plus est, le principe même de la création et du lancement de tendances consiste à aller dans des zones non explorées. S'appuyer sur trop d'IA pour détecter des tendances pourrait amplifier « l'effet du rétroviseur ». Des outils d'IA ou de représentation des tendances peuvent servir à comprendre l'environnement dans lequel évolue une marque. C'est ce que propose **Heuritech** qui analyse les tendances de la mode dans des bases d'images hétérogènes, avec les couleurs et motifs qui plaisent.

L'IA peut aussi servir à créer de nouveaux parfums. Plusieurs projets vont dans cette direction comme celui d'**IBM Research** et **Symrise**, un producteur de parfums qui fournit Estee Lauder et L Brands. Leur AI dénommée Philyra s'appuie sur du machine-learning pour exploiter la base de données de Symrise qui comprend 1,7 millions de formules et identifie les "oceans bleus" de formules non testées ou non utilisées (vidéo). Le système suggère alors des nouvelles combinaisons d'essences et de parfum qui pourraient plaire à certains segments de clientèle[1550].

En 2019, **Google Brain** publiait un outil de détermination de l'odeur d'une molécule en fonction de sa composition, entraîné avec une base de 5000 molécules labellisées[1551].

Edited (2009, UK, $6M) est un outil d'aide à la décision de création de gammes. Il permet aussi de réaliser une analyse concurrentielle, via l'analyse de défilés de modes et de sites de e-commerce.

Production

Le marché du textile peut aussi faire appel à des robots, comme ceux de **Softwear Automation** (2007, USA, $7,5M) qui sont capables de fabriquer des chaussures, des t-shirts et des matelas (vidéo). Ce sont surtout des automates qui répètent les mêmes gestes à longueur de journée mais ils disposent tout de même de systèmes de vision artificielle pour gérer leur production.

Personnalisation

Dans l'habillement, la personnalisation des produits via des scanners 3D, de la reconnaissance d'images prises avec des systèmes de captation 3D en boutique ou des smartphones permettent de déterminer les mesures exactes et/ou la corpulence du client ou de la cliente. Les clients se voient alors proposer les bonnes tailles de vêtement.

Ce genre de méthode intervient aussi de plus en plus dans le calibrage des montures de lunettes aussi bien pour la vente en opticien qu'en ligne. Là encore, avec de la reconnaissance visage et des éléments clés dans les visages.

La marque **Stitch Fix** utilise un « styliste digital » paramétré par les clients avec leurs goûts et préférences vestimentaires, leur mode de vie, loisirs, choix de looks. Le tout récupéré notamment via leur compte Pinterest. L'IA propose ensuite des choix vestimentaires correspondant à ce profil.

[1549] C'est le propos de l'article Intelligence artificielle et luxe : l'alliance des possibles de Laura Perrard, mai 2018.

[1550] Voir Breaking new fragrance ground with artificial intelligence (AI): IBM Research and Symrise are working together, octobre 2018. La solution rappelled aux industries traditionnelles qu'il est bon d'avoir un système d'information qui capte bien l'historique métier de la société.

[1551] Voir Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules, octobre 2019 (18 pages).

La marque **Eison Triple Thread** produit des costumes sur mesure avec sa solution FITS en s'appuyant pour sa part sur les goûts musicaux du client qui sont détectés sur sur compte Spotify et associés à un questionnaire de profiling voisin de celui de Stitch Fix[1552]. Le système génère alors des styles de costumes alignés sur ces goûts. In fine, c'est toujours le client qui choisit.

En 2019, **Facebook** présentait son réseau de neurones Fashion++ qui permet de choisir son accoutrement pour être « fashion ». C'est pour l'instant un projet de recherche. Comme il se doit, il est entraîné par force brute avec des milliers de combinaisons de vêtements jugées correctes pour le look. Il propose alors de modifier une combinaison existante pour la rendre plus fashion[1553].

Il est même possible d'utiliser des réseaux de neurones génératifs de type GAN pour animer des vêtements arbitraires sur des mannequins modèles (vidéo).

Expérience client

L'IA peut servir à trouver des produits similaires dans les catalogues en ligne et, dans la lignée, identifier des contrefaçons ou produits gris, au niveau des produits eux-mêmes comme à celui du packaging.

Farfetch (2008, UK, $705M) est un site de e-commerce et une place de marché dans la mode et le fashion. Il s'appuie sur de l'IA pour améliorer les parcours clients avec la plateforme de **Certona** (2004, USA, $37M).

On peut peut évaluer l'émotion « extérieure » des clients en les détectant sur les visages captés par des caméras dans les boutiques ou sur mobiles et laptops. C'est ce que proposent des sociétés comme **Affectiva** ou **Datakalab** qui analysent le visage d'une ou plusieurs personnes pour capter différents types d'émotions affichées. Cela peut être exploité avec de la vidéo surveillance de boutique pour évaluer la performance d'une devanture ou de contenus d'affichage dynamique.

Contrefaçons

La détection de contrefaçons est un objectif clé de nombreuses marques de luxe, qu'il s'agisse des sacs de Louis Vuitton, des parfums ou dans la joaillerie. Celle-ci concerne aussi bien les produits que les packagings.

En plus de **Data & Data**, déjà évoqué dans le reste du document, nous avons aussi **Entrupy** (2012, USA, $2,6M) approche la recherche de contrefaçons de manière originale avec une petite caméra portative photo permettant de prendre des photos en mode macro des objets et d'analyser leur matière avec une IA. La solution serait fiable à 99,1%, ce qui est plausible si la base d'entraînement est bien alimentée. Et pour cause, puisqu'elle contient plus de 30 millions d'images (vidéo). Une technique équivalente serait capable de détecter le coup de pinceau de peintres dans 80% des cas et d'identifier des faux[1554].

Cypheme (2015, France/USA, $1,3M) est autre une startup dans ce secteur, qui propose une application mobile de détection de produits contrefaits, s'appuyant sur un algorithme de machine learning appliqué à la qualification d'images, associé à une étiquette spéciale. C'est une sorte de Shazam de la contrefaçon.

[1552] Voir Fashionably AI - Eison Triple Thread reimagines suit-making with music and artificial intelligence, de Henry Pickavet, août 2018.

[1553] Voir Fashion++: Minimal Edits for Outfit Improvement, septembre 2019 (10 pages).

[1554] Voir AI Used To Spot Art Forgeries By Looking At Brushstrokes de Tyler Lee, novembre 2017.

Services et conseil

Les cabinets de conseil, d'études de marché, analystes, services et autres consultants indépendants sont les premiers à secouer le marché dès qu'une nouvelle technologique apparaît.

Ils pratiquent à outrance le marketing de la peur, expliquant aux entreprises qu'elles doivent à tout prix adopter ces nouvelles technologies sous peine de se faire disrupter ou uberiser en cas d'immobilisme technologique[1555].

Ces métiers représentent plus de 450 000 emplois en France rien que dans le conseil et la réalisation de projets informatiques !

Mais ces cordonniers sont-ils bien chaussés ? Est-ce que ces « travailleurs du savoir » sont bien outillés avec l'IA pour être performants et pertinents ? Au nez, pas vraiment. Les méthodes de travail des consultants et des analystes n'ont pas beaucoup évolué, avec ou sans IA. Mais il y a du potentiel et les idées ne manquent pourtant pas. C'est un monde d'abondance de *couldashouldawoulda*[1556].

Pourquoi en est-on à ce point ? La faible automatisation d'un métier a plusieurs origines : sa grande fragmentation organisationnelle, la diversité des tâches et des situations, la déstructuration des données utilisées, le fait que l'Internet soit un gros foutoir mal organisé et moins ouvert qu'il n'y parait, l'absence de standards et les limites actuelles des techniques logicielles et de l'IA.

Qui plus est, le marché de l'outillage du conseil est aussi un marché relativement limité. Lorsque l'on additionne la fragmentation des besoins et celle du marché, cela donne un marché faiblement innovant côté outillage. Sauf lorsque ce marché peut réutiliser des outils conçus pour des marchés plus grands.

Passons maintenant en revue quelques-unes de ces applications potentielles de l'IA dans différents métiers du conseil en passant par le conseil en stratégie et en management, la gestion de projets, les études de marché et les projets informatiques.

Conseil en stratégie, innovation et management

Le conseil en stratégie, innovation et management démarre souvent par une phase de découverte de l'entreprise cliente avec un grand nombre d'entretiens avec ses collaborateurs et même avec d'autres entreprises. Cette tâche peut être facilement fastidieuse car il faut généralement conserver une trace écrite de ces entretiens, puis en faire une synthèse qualitative et quantitative[1557].

C'est là que peuvent être exploités des outils de transcription automatique des entretiens s'ils sont enregistrés. Le speech-to-text est une technique encore imparfaite mais elle peut permettre de gagner du temps. Elle est même intégrée dans certains outils bureautiques comme **OneNote** de Microsoft.

La synthèse automatique peut être réalisée par l'extraction automatique d'éléments clés dans le texte. Cela peut concerner des expressions ou des mots qui sortent de l'habitude, ou par la détection de sentiments dans les textes.

[1555] Le message le plus classique consiste à mettre en regard la valorisation boursière de sociétés du numérique (Uber, AirBnb, Facebook, Google) et celle d'entreprises traditionnelles (automobile, BTP, hôtellerie), en négligeant le chiffre d'affaires et la profitabilité. Bref, cela consiste à comparer des choux et des carottes. Si toutes les entreprises étaient des « pure players » numériques, on serait bien avancé et incapables de se loger, de s'alimenter, de se chauffer, d'accéder à de l'eau potable et de se déplacer. La comparaison n'est pas la bonne. Toutes les entreprises qui sont dans les biens de consommation ou infrastructures matérielles n'ont pas forcément vocation à quitter ces métiers, même si cela peut parfois arriver, comme la manière dont le groupe Accor Hôtels a séparé l'activité de l'hôtellerie de celle de l'immobilier associé. Bref, la « transformation digitale » ne doit pas prendre comme repère les pure players de l'Internet.

[1556] Comme évoqué dans L'intelligence artificielle au service des cabinets de conseil, par Matthieu Gufflet, CEO d'Epsa Groupe et Thomas Bourgeois, CEO de Dhatim, juillet 2018, qui fait la promotion de la notion de « consultant augmenté ».

[1557] Voir Le conseil de l'innovation mise sur l'intelligence artificielle, 2018.

Des statistiques des mots clés employés peuvent aussi être réalisées. C'est ce qu'avait fait la société **Proxem** pour l'équipe En Marche en 2016 en amont du lancement de la campagne présidentielle d'Emmanuel Macron.

Ces analyses recouvrent parfois une analyse de la concurrence qui passe par des entretiens et l'exploitation de sources ouvertes ou fermées qu'il faut consolider. Il est notamment critique de récupérer des indicateurs économiques quantitatifs sur ces concurrents. L'IA intervient peut dans ce processus à ce jour.

Dans les phases d'idéation, des de représentation des connaissances peuvent être exploités avec des arbres de décision, des métaplans virtuels créés sur tableaux de bord interactif. Mais là encore, sans forcément faire appel à de l'IA. Elle pourrait cependant intervenir pour réaliser automatiquement des regroupements syntactiques ou thématiques de diagnostics ou de propositions.

Suit en général la production de rapports et de présentation. Leur efficacité peut être améliorée avec des outils d'analyse du langage : corrections, calculs d'indices de lisibilité et analyse d'efficacité d'argumentaires. Il est bien plus facile d'automatiser les prestations de services sur des métiers opérant sur des données quantitatives comme la finance, la comptabilité, les achats, les facturations, les audits et le marketing.

Gestion de projets

C'est une activité courante dans les cabinets de conseil et les entreprises de services numérique (ESN). Les outils traditionnels utilisés servent à la gestion des plannings et des ressources. Ce sont des gestionnaires de projets du marché qui peuvent intégrer des briques d'IA selon les cas de figure.

Ils peuvent intégrer des fonctions prédictives et identifier à l'avance les zones de congestion ou de retards potentiels des projets, histoire de les prévenir en allouant les ressources de manière appropriée. La question clé étant de disposer de données d'entraînement suffisantes pour que les IA se débrouillent.

Lili (2016, France) capte l'ensemble des données des projets dont les communications écrites entre intervenants pour prévenir ou gérer les litiges.

Des solutions à base d'IA peuvent aussi être mises en place pour le staffing des projets, pour trouver les bonnes compétences, vérifier la compatibilité humaine des équipes, composantes que nous avons partiellement évoquées dans la rubrique sur les ressources humaines.

Etudes de marché et sondages

Les études de marchés et sondages génèrent de gros volumes de données qui se prêtent à des analyses à base d'IA. C'est le métier de sociétés telles que Médiamétrie, BVA, Ipsos ou l'IFOP mais aussi le Gartner, Forrester ou IDC.

Les enquêtes terrains et sondages peuvent éventuellement faire appel à des chatbots textuels ou vocaux pour le questionnement des sondés. Cela a du sens économique mais peu biaiser les réponses générées.

Dans les baromètres d'audience de médias chauds comme la TV, des systèmes d'analyse du signal audio qui sort de la TV permettent d'identifier les programmes visualisés. C'est de l'IA ! D'autres solutions détectent automatiquement des logos dans les vidéos, pour évaluer la présence de marques et annonceurs dans les contenus à la TV tout comme dans les médias en ligne. L'analyse de sons et d'images en tout genre avec du deep learning devient un moyen clé d'auditer l'exposition médiatique et conversationnelle des marques.

De leur côté, les études de marché qualitatives impliquent la recherche, la compilation et la mise en forme de données hétérogènes issues de nombreuses sources. Cela comprend des recherches bibliographiques sur Internet, l'usage de moteurs de recherche et de bases ouvertes ou payantes spécialisées.

Il y a encore beaucoup à faire pour automatiser la curation d'information. Les méthodes utilisées restent en général assez traditionnelles avec diverses astuces d'utilisation des moteurs de recherche, la recherche par mots clés dans des flux RSS, des systèmes d'alertes.

Ces techniques pourraient exploiter de l'IA pour faire des analyses syntaxiques, créer automatiquement des ontologies de domaines d'activité.

Quid d'un **Rapport du CES de Las Vegas** qui serait généré automatiquement par une IA, ou même cet ebook ? Une bonne part de mon travail de rédaction relève d'un processus itératif. Découverte de nouveaux produits, recherche de leur description la plus complète et factuelle, recherche d'illustrations photos et vidéo, puis synthèse. Théoriquement, tout cela pourrait bien être partiellement automatisé. Il en va de même des comparatifs qui pourraient aussi l'être, par exemple entre tous les smartphones ou appareils photos du marché. Pourtant, ce qui est théoriquement possible ne l'est pas pratiquement ou économiquement car peu de professionnels ont ce genre de besoins. Ce n'est pas scalable !

On peut aussi faire appel à diverses bases de données d'entreprises et de startups qui consolident les informations publiques à leur sujet comme **Crunchbase**, **Skopai**[1558] ou **U-Change**[1559]. L'analyse de vidéo et leur transcription automatique bien mise en forme pourrait aussi aider à trouver les informations pertinentes. La recherche d'information pourrait aussi bénéficier de l'existence d'outils d'élagage de textes, supprimant la rhétorique, les superlatifs, termes vagues et autres discours stéréotypés.

Ce genre d'activité implique aussi la consolidation de données chiffrées pouvant faire appel à des outils d'extraction de données situées dans des textes. Je passe généralement pas mal de temps à générer des tableaux contenant les grands indicateurs économiques de quelques dizaines de sociétés du monde du numérique. Cette compilation est encore bien trop manuelle. Et ne parlons pas des sociétés non côtées pour lesquelles les informations recherchées ne sont pas dans le domaine public !

Enfin, il faut produire des visuels et notamment des charts et autres business analytics. Et puis surtout, il faut trouver des moyens de vulgariser certains contenus techniques. L'utilisation d'analogies ou d'exemples n'est pas encore à la portée des IA de traitement du langage ou de génération de résumés.

Des modèles mathématiques plus sophistiqués à base de machine learning pourraient aussi être développés pour croiser plusieurs études de marché, identifier les variations et écarts types, normaliser les études consolidées en fonction de leurs descriptifs qui sont souvent hétérogènes. Comme par exemple la comparaison des prévisions sur le pourcentage des emplois qui seront détruits et créés par l'IA dans les années à venir. Des outils de normalisation des études en termes de méthodologie, de périmètre, de champs et de timing seraient les bienvenus.

[1558] **Skopai** (2017, France) utlise divers outils d'IA provenant notamment du laboratoire LIG de Grenoble (une UMR CNRS - Grenoble INP – Inria et Université Grenoble Alpes) pour générer une base de données de startups et PME technologiques alimentée par les données ouvertes du web. Cela peut aussi servir aux investisseurs qui cherchent à déterminer un panorama concurrentiel d'une startup.

[1559] **U-Change** (2018, France) est un autre projet de génération de base de données de startups exploitant des données ouvertes du web. La société a d'autres activités tournées autour du respect du RGPD et à la transformation digitale des PME.

L'une des raisons pour lesquelles cette consolidation est difficile est la faible structuration des données publiées sur Internet. L'adoption des formats XML et RDF promus sur Internet au début des années 2000 a été faible[1560]. Les initiatives de publication d'open data se multiplient mais leur exploitation reste un processus très manuel.

Les études de marché pourraient exploiter diverses sources de données que ce soit par exemple dans Google Trends, LinkedIn et d'autres réseaux sociaux. Cela permet d'analyser les mouvements entre entreprises, les évolutions des compétences affichées dans les CV, les postes ouverts, la circulation des talents. Ces outils ne reposent pas encore assez sur de l'IA. Ils pourraient y faire appel au minimum dans l'interprétation de données non structurées.

D'autres sources permettent aux entreprises de mieux connaître leurs concurrents. Leur CA, leurs clients, les gains et pertes de projets, la visivilité de leurs produits, les évolutions des équipes, des méthodes marketing, les verbatim marketing, la satisfaction client, l'analyse de sentiments, celle des signaux faibles, les analyses temporelles, l'identification de leurs influenceurs. Voilà de quoi probablement alimenter quelques briques d'IA de traitement du langage et des données.

On rêve de moteurs de recherche d'un nouveau genre qui seraient capables d'interpréter nos questions et d'aller chercher tout seul les éléments de réponse sur Internet, les intégrer et les mettre en forme. C'est technique possible. Dans nombre de cas, il ne s'agit pas de créer des AGI (intelligences artificielles générales). Reste à le faire ! Qui se lance ?

Services informatiques

Terminons avec les services informatiques. Chez eux, qu'est-ce qui est répétable et automatisable ? Nombre de processus qui ne relèvent pas forcément de l'IA peuvent être automatisés avec de simples macros.

Dans l'IT, l'informatique d'infrastructures, les outils de monitoring commencent à exploiter le machine learning pour détecter les bizarreries et les tendances. C'est ce que nous avons vu dans le cadre des AIops. La RPA permet d'automatiser de nombreux processus de l'IT.

Est-ce que les développeurs seront remplacés par des IA ? Certaines actualités récentes le laissent croire mais c'est un mythe. Il me rappelle le débat sur les langages de quatrième génération dans les années 1980. Leur promesse n'a jamais été véritablement tenue même si l'on a pu bénéficier de progrès via la prolifération de « frameworks » associés aux grands langages de programmation du marché (Java, JavaScript, PHP, Python…). Mais on a toujours autant besoin de développeurs. Tout simplement parce que la quantité de logiciels développés à augmenté, Internet et mobile obligent.

Education

Le secteur de l'éducation est depuis longtemps un grand champ de promesses pour les usages de l'intelligence artificielle. Les applications pressenties touchent souvent à la personnalisation de l'enseignement à distance via des agents conversationnels intelligents capables de suivre et accompagner pas à pas les élèves dans leur progression[1561]. On en est encore au stade des expérimentations. La littérature sur le sujet est pour l'instant assez vague[1562].

[1560] Je m'en suis une fois de plus rendu compte en développant une petite macro VBA pour la création de cet ebook. Celle-ci a pour objet de transformer une URL en texte clair avec « Voir (nom de l'article + hyperlien), auteurs, date ». Et bien, les tags META d'HTML sont utilisés de manière complètement désordonnée dans le web d'aujourd'hui. C'est un énorme bazar. Et pourtant, ce ne serait pas de la science-fiction si les éditeurs de CMS et de sites web adoptaient des schémas standardisés pour décrire les principales métadonnées de leurs pages et sites web ! Cette anecdote illustre le décalage qui peut exister entre les deep techs, l'IA, le bon sens humain et la collaboration qui sont toujours aussi déficients.

[1561] La version la plus extrême de ces méthodes est le « neurohacking » des élèves qui consiste à décortiquer les processus de mémorisation des connaissances et à hacker le cerveau pour le faire fonctionner de manière optimale. Cela va jusqu'à faire des IRM fonctionnelles pendant des phases d'apprentissage pour identifier les parties du cerveau qui sont activées.

[1562] En voici un exemple: The Future of AI and Education, mai 2018, qui n'est pas bien dense en études de cas !

Ces solutions sont difficiles à mettre au point et à déployer à grande échelle. L'éducation est un de ces secteurs d'activité où la fragmentation ralentit l'automatisation. Ici, il s'agit de la fragmentation des pratiques d'enseignement, des matières, des contenus, des méthodes et typologies d'élèves, multipliés par les variantes linguistiques. Comme les logiciels en général, l'IA est pertinente lorsqu'elle génère des économies d'échelle. Et ici, ce n'est pas évident.

Dans Quelle place pour les robots dans le tutorat à distance ?, octobre 2016, Jacques Rodet pose de bonnes questions sur les usages du numérique et de l'IA dans l'enseignement.

Le tableau *ci-contre* décrit les tâches dans l'enseignement qui peuvent être automatisées et celles qui demanderaient encore des enseignants.

Plus de la moitié des tâches requièrent ces derniers dans le tableau.

Fonctions	Plan cognitif	Plan Socio-affectif	Plan motivationnel	Plan métacognitif
Accueil et orientation	Informer sur le dispositif de formation	Initier la construction d'un sentiment d'appartenance	Faire émerger les objectifs personnels de l'apprenant	Inciter l'apprenant à faire le point sur ses stratégies cognitives
Organisation	Présenter les méthodologies appropriées	Réguler la dynamique de groupe	Accompagner le processus d'autonomie	Faciliter la planification de l'apprentissage
Pédagogie	Apporter des réponses ou les susciter. Remédier	Faciliter la collaboration des apprenants	Proposer des activités signifiantes	Susciter l'expression critique sur le dispositif
Socio-affectif Motivation	Personnaliser le soutien à l'apprentissage	Rompre l'isolement de l'apprenant	Lutter contre l'abandon	Faire prendre conscience de ses habiletés à collaborer
Technique	Aider à maitriser l'environnement d'apprentissage	Susciter l'entraide technique entre apprenants	Encourager l'utilisation des outils	Susciter la prise de distance réflexive sur les usages des outils
Métacognition	Faire conscientiser ses préférences cognitives	Faciliter la prise de conscience des états affectifs / tâches	Faire identifier les motivations intrinsèques	Inciter l'apprenant à apprendre à apprendre
Evaluation	Annoncer clairement les critères d'évaluation	Produire des rétroactions à portée formative	Encourager et féliciter	Aider à s'autoévaluer

En gris les interventions pouvant être confiées à un robot ayant été préalablement formé
En bleu, les interventions partagées par un robot et les tuteurs humains
En vert, les interventions ne pouvant être confiées qu'à des tuteurs humains

Applications

Voici quelques-unes des applications potentielles de l'IA dans l'éducation qui sont couramment citées dans la littérature sur le sujet voire expérimentées ou intégrées dans des offres commerciales[1563] :

Notation automatique des élèves, ce qui n'aurait de sens que pour noter la partie qualitative de QCMs. Pour ce qui est de la notation de copies manuscrites d'élèves, les IA de reconnaissance d'écriture ne sont pas encore assez au point.

Accompagnement d'élèves à déficits cognitifs, qui peut exploiter diverses techniques de captation et de suivi de l'attention[1564]. Cela comprend l'usage de robots accompagnant les enfants dans certains processus d'apprentissage, notamment à destination d'enfants aspergers ou autistes.

C'est l'application du petit robot de **Leka** ainsi que le **Nao** de Softbank Robotics, anciennement Aldebaran Robotics.

[1563] Cet inventaire est inspiré de 7 roles for artificial intelligence in education de Matthew Lynchmay, mai 2018 et de Education and AI from Intelligence Unleashed, An argument for AI in Education de Rose Luckin, UCL Knowledge Lab, University College London.

[1564] Voir How AI is changing special education, juin 2017.

Support des élèves via un chatbot, ce qui suppose un environnement pédagogique bien formalisé dans le chatbot en question[1565]. Le chatbot devrait aussi avoir accès à l'historique personnel de l'élève. Bref, savoir tout sur lui de sa scolarité et sur ses forces et faiblesses. Dans un monde parallèle ! Mais sur des tâches élémentaires, ce genre d'outil allégerait la tâche des enseignants, leur permettant de mieux jouer le rôle de mentor. Ces outils permettraient aussi d'accompagner les élèves hors des classes et dans les MOOCs pour personnaliser l'enseignement, l'assimilation des compétences et la réussite des tests. Attention cependant : diverses études récentes ont montré que les élèves faisant de l'apprentissage uniquement via des MOOCs obtenaient de moins bons résultats que les élèves passant par des cours en présentiel. L'IA peut-elle arriver à la rescousse ?

Apprentissage de l'écriture avec l'analyse de l'écriture manuscrite automatique qui fonctionne de manière limitée, et plus efficacement sur la détection de lettres individuelles. Des outils de ce type sont proposés sur tablettes dotées de stylets.

Outils d'accès à la connaissance. L'éducation, surtout dans le secondaire et le supérieur peut bénéficier de l'usage de moteurs de recherche plus intelligents et d'outils voisins de ceux que j'imaginais dans la partie précédente sur les services et le conseil. Ils peuvent être complétés d'autres outils, évoqués dans la rubrique médias et qui permettent de détecter des fausses nouvelles ou informations (« fake news ») aussi bien sur des aspects qualitatifs que quantitatifs.

Accompagnement d'une classe, permettant aux enseignants d'identifier les forces et faiblesses d'un groupe d'élève, les parties bien assimilées ou pas d'un cours, le niveau de leur attention. Des techniques utilisant des caméras et autres capteurs permettent déjà de capter l'attention générale d'une audience et de segmenter les groupes d'individus par comportement. Cela permet aussi d'ajuster le contenu de cours car l'attention dépend aussi des pratiques de l'enseignant. Cela fait partie du périmètre de sociétés américaines telles que **Content Technologies** (2005, USA), qui cependant ne cible pas que le secteur de l'éducation, modèle économique oblige[1566] et **Carnegie Learning** (1997, USA, acquis par Apollo Education Group, $14M) qui est focalisé sur l'enseignement des mathématiques dans le secondaire et le supérieur[1567].

A partir des résultats de tests, des solutions à base de machine learning permettent une classification automatique des élèves en groupes homogènes.

Détection de fraude et de plagiat, une application qui n'est pas souvent évoquée et qui porterait sur les épreuves de concours et sur les thèses, en disposant de versions numériques des textes des élèves, la reconnaissance d'écriture à grande échelle n'étant pas suffisamment fiable.

Serious games. Dans certains contextes, les jeux peuvent être mis en place pour certains apprentissages, aussi bien de l'écriture que des mathématiques, de physique ou de sciences de la vie. L'IA peut en théorie aider à les conçevoir et à les rendre personnalisables en fonction de chaque enfant. Se posera cependant à chaque fois la question de l'apprentissage de l'IA elle-même et des données qui l'alimentent. S'il s'agit de règles et de faits, on retombe dans des moteurs de règles traditionnels. S'il s'agit d'exploiter des données de jeux d'autres élèves pour faire de l'apprentissage supervisé ou non d'une IA, alors le volume de données nécessaire pourrait dépasser l'entendement. A la fin, l'usage d'IA dans ce créneau dépendra de l'équation économique de la formation. Donc, de son audience. Ce qui favorisera les langues dominantes dans les pays développés, l'anglais en premier.

[1565] Voir A college professor used an AI teaching assistant for months, but his students didn't notice, 2016.

[1566] Content Technologies cible aussi les marchés de la santé, du secteur public et de la finance.

[1567] Citées dans How Is AI Used In Education - Real World Examples Of Today And A Peek Into The Future, de Bernard Marr en juillet 2018.

Aides sur le développement kinétique avec des solutions divers exploitant plus ou moins d'IA dans les sports, la danse, les apprentissages de mouvements dans l'espace en général. L'analyse visuelle des corps permet de diagnostiquer les points d'amélioration. Nous avons aussi vu qu'il existait déjà des générateurs de mouvements synthétiques exploitant des réseaux génératifs. En pratique, pour créer des vidéos avec les visages et le corps des enfants se déplacement sur des mouvements réalisés par d'autres enfants.

Analyse de l'attention des élèves et de la performance des enseignants. Là encore, des outils vidéo de détection des émotions et de l'attention pourraient permettre de suivre les élèves inattentifs ainsi qu'évaluer les enseignants qui n'arrivent pas à capter l'attention de leurs élèves.

Ce sont cependant des domaines où il faut se méfier du solutionnisme technologique selon lequel on peut résoudre tous les problèmes avec de la technologie. Ici, la méthode est contestable et probablement difficile à justifier économiquement pour l'équipement de toutes les salles de classes. Mais le monde est grand et cela sera sûrement expérimenté quelque part si ce n'est pas déjà fait.

Education à l'IA. Tous les outils cités précédemment sont de l'IA appliquée à l'éducation. Reste à s'intéresser à l'enseignement de l'IA elle-même. Il est probable qu'il se positionnera dans la lignée des enseignements du numérique avec d'un côté des enseignements sur les usages des outils exploitant de l'IA, et de l'autre, à la création de ces outils pour les futurs professionnels du secteur. Cela passe par l'explication des concepts de base de l'IA, la description des applications dans le raisonnement automatique, le traitement des données, celui des langages et de la vision, puis des systèmes qui les embarquent (mobiles, robots, véhicules autonomes, chatbots, agents vocaux…). Pour les techniciens et ingénieurs, il faudra passer par des mathématiques, des concepts nombreux autour des réseaux de neurones, des outils de développement, de la data science, de l'architecture de systèmes et de l'intégration.

Automatisation des processus administratifs. Et là, il y a du boulot ! La plupart des établissements sont loins d'avoir fait leur « transformation digitale ». J'en fais moi-même l'expérience chaque année lorsque je dois remplir n fois un dossier de vacataire pour mes conférences dans l'enseignement supérieur. Les formulaires sont remplis sur papier. La modernité consiste à annoter un PDF qui peut être envoyé par email. Il n'y a pas de formulaire en ligne. Il faut remplir le même formulaire chaque année et fournir à chaque fois les mêmes pièces administratives. Bref, c'est l'âge de la pierre. Et vous croyez que c'est là que l'on va développer des solutions d'IA ? Je crains fort que cela ne soit pas le cas, tout du moins en France. Tout du moins pour les processus concernant les enseignants. Pour les élèves, pourquoi pas, là où ils sont nombreux et dans les grands établissements et grandes filières. Au passage, la majorité des processus administratifs ont besoin d'applications numériques traditionnelles et l'IA est superfétatoire dans un premier temps.

Recrutement. L'IA est mise en œuvre aux USA dans les universités pour les aider à recruter les meilleurs étudiants comme chez **Plexuss** (2014, USA, $3,1M) ou, au contraire, pour aider ces derniers à trouver la meilleure université avec **Admitster** (2014, USA).

Sécurité d'accès. Elle exploite la vidéo surveillance et la détection des visages. C'est une application que l'on ne voit pas encore en France, heureusement, mais qui se développe en Chine[1568].

Startups

J'ai identifié quelques startups dans le monde qui s'attaquent aux usages de l'IA dans l'éducation et de manière très différente. Comme partout dans ce document, cet inventaire est bien plus illustratif qu'exhaustif.

[1568] Voir Artificial intelligence is watching China's students but how well can it really see?, par Echo Xie, septembre 2019.

- **AiEducate** (UK) a créé DALI (Dialog-based Artificial Learning Intelligence), un chatbot de tutorat. Le chabot joue deux rôles : il gère des conversations en mode questions/réponses classiques et enseigne. Ils sont partenaires d'IBM Watson, ayant participé au concours Xprize d'IA organisé par ce dernier. Il n'est pas évident que ce projet ait une vie commerciale.
- **AskMyClass** (USA) s'intègre avec Amazon Alexa et Google Home, pour faire faire des exercices aux élèves du primaire, dans l'apprentissage du vocabulaire. Il permet aussi aux enseignants de prendre des notes. Le chatbot est censé faire gagner plus d'une heure de travail aux enseignants par semaine. La solution est commercialisée en mode freemium. C'est évidemment adapté au marché US ou tout du moins anglophone.
- **Gradescope** (2014, USA, $5,3M) divise par deux le temps que les enseignants passent pour noter les copies des élèves d'examens passés en ligne. Cela passe par de la reconnaissance d'images et fonctionne visiblement pour les sciences « dures » (maths, physique, chimie, biologie, informatique). Le système regroupe les réponses par catégories ce qui permet de noter une seule fois chaque réponse différente.
- **Skillogs** (2015, France) a développé Acarya, du Sanskrit "Professeur", un environnement d'apprentissage qui permet un apprentissage adapté à la vitesse de chaque élève, selon leur niveau (vidéo). L'IA utilisée ? Bien, elle n'est pas précisée, comme c'est souvent le cas. On ne sait en particulier pas identifier les données d'entraînement de cette IA. Il s'agit d'une banque de données de compétences, d'analyse des forces et les faiblesses des élèves. C'est encore une solution qui s'appuie sur IBM Watson.
- **Domoscio** (2013, France, $569K) propose un « *outil d'ancrage adaptatif mémoriel* » fondé sur l'IA et le machine learning (pourquoi cette distinction ? L'IA pour le traitement du langage et le ML pour l'analyse des données ?). Il analyse le niveau et la manière d'apprendre des élèves pour leur proposer le contenu adapté et au bon moment, histoire de favoriser la mémorisation. C'est visiblement plutôt utilisé pour la formation continue, dans de grandes entreprises (SNCF, Banque de France, Bouygues Télécom) et dans l'enseignement supérieur (Sciences Po, Universités).
- **Soul Machines** (Nouvelle-Zélande, $7,5M) a développé des avatars très réalistes, qui sont intégrés dans la solution de formation sur les énergies renouvelables de **Vector** (vidéo). On n'est pas très loin du syndrome de la vallée de l'étrange, ce sentiment curieux et dérangeant que l'on ressent (en Occident, pas en Asie) lorsque l'on interagit avec une machine qui adopte l'apparence humaine.
- **Groupe Bizness** (2007, France) utilise #skillgym pour la formation sur-mesure dans la vente, le management, la transformation digitale (?), les relations et la satisfaction client. Ils testent la réaction émotionnelle des élèves dans des mises en situation (via une tablette et leur webcam) avec des vidéos interactives jouées par des acteurs. Leurs clients sont dans la banque, les assurances et l'automobile. (vidéo)[1569].

[1569] L'usage de l'IA dans la formation continue fait écho à l'impact de l'IA sur les métiers. The role of education in AI (and vice versa), avril 2018, évoque les compétences à acquérir dans un monde professionnel entouré d'IA. Le gagnant est le développement de la créativit et des compétences sociales et de perception. Bref, des *soft skills*.

- **Snapask** (2015, Hong Kong, $21,8M) a développé une application qui permet de mettre en relation des élèves et des tuteurs pour des séances de coaching thématiques. Les élèves prennent une photo du problème à résoudre et le décrivent avec quelques mots. L'application se charge ensuite de la mise en relation avec le tuteur le plus approprié. Ils affichent en avoir déjà 120 000. Ils sont rémunérés à hauteur de $1 par question traitée. La startup n'est pas encore véritablement internationalisée. C'est une idée parmi d'autres.
- **Squirrel AI Learning** (2014, Chine) a dévelopé une solution d'apprentissage adaptatif à base d'IA ciblant le marché chinois et les élèves de l'enseignement primaire[1570]. La solution repose sur des tests d'évaluation et de diagnostic précis de lacunes de compétences des élèves, notamment dans le raisonnement logique et la gestion de graphes de connaissances. Elle propose alors une formation adaptée à la situation avec des cours en ligne et des exercices personnalisés. On imagine bien que la pertinence de la solution dépend étroitement de la volumétrie des données d'entraînement qui ont alimenté ses modèles. On supposera que ce volume est élevé puisque l'on est en Chine.

  Ils ont recruté pour ce faire un certain Tom Mitchell, un spécialiste de l'IA provenant de Carnegie Mellon, avec qui ils ont monté un partenariat de recherche. Ils sont aussi partenaire du laboratoire privé SRI aux USA.
- **RoboTutor** (2017, USA) est un logiciel open source d'apprentissage de l'écriture et de la lecture pour les enfants non scolarisés de 7 à 10 ans utilisant un système de reconnaissance automatique de l'écriture et de la parole. C'est un projet de l'Université Carnegie-Mellon ayant été finaliste du concours X Prize for Education.

Services publics

Au même titre que le numérique de manière générale, l'IA peut être mise en œuvre quasiment partout dans les processus des services publics. Dès lors qu'il y a des processus et des données ou de l'information, l'IA peut jouer un rôle. Le cas le plus classique est l'amélioration de la relation avec les citoyens ou clients[1571]. Cela peut notamment passer par l'usage de chatbots.

Les données chiffrées de nombreuses activités peuvent servir à réaliser des prévisions, comme pour faire des simulations de rentrées ou de dépenses fiscales.

Les applications sont plus nombreuses dans le contexte de la ville intelligente pour prédire l'usage des infrastructures, optimiser la consommation d'énergie ou gérer la sécurité par la vidéo-surveillance.

Police et justice

L'IA peut aussi servir pour la police et la justice. La police de certaines villes comme **Londres** utilise l'analyse des images de caméras de surveillance dans les gares pour détecter les anomalies. Cette surveillance est particulièrement utile pour surveiller les grands événements sportifs ou autres pour détecter d'éventuelles menaces le plus en amont possible.

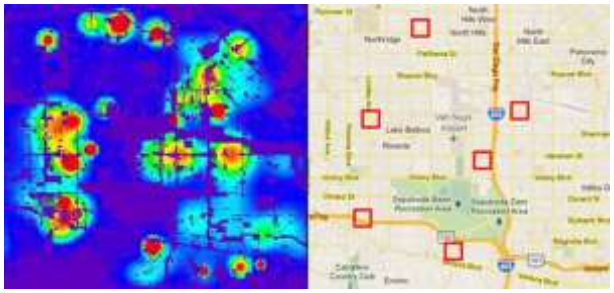
[1570] Voir Forbes Insights: How AI Can Realize The Promise Of Adaptive Education par Derek Haoyang Li, 2020.

[1571] Voir Artificial Intelligence for Citizen Services and Government, août 2017 (19 pages).

La reconnaissance faciale permet à la fois de trouver des criminels en fuite, de les pister, mais aussi de retrouver des personnes portées disparues. C'est le business de la controversée startup **Clearview** (2017, USA, $8,6M) [1572]. À Dubaï, des robots à roulettes et écran tactiles permettent de reporter des incidents et de transmettre des images en direct. Il existe même des robots de déminage, en général télécommandés.

La police de la ville de **Durham**, au Royaume-Uni expérimentait en 2017 une application développée par l'Université de Cambridge servant à classifier les suspects arrêtés pour évaluer leur niveau de risque, exploitant quatre années d'archives d'arrestations. Le système est dénommé HART pour **Harm Assessment Risk Tool**[1573].

La startup **Predpol** (2012, USA, $3,7M) a déjà une offre avec ce même positionnement pour prédire les lieux et les fenêtres temporelles où des crimes pourraient être commis. La solution est exploitée pour planifier les patrouilles de la police locale. Avec le risque de générer des biais liés aux données d'entraînement, focalisant les forces de l'ordre sur les personnes de couleur, dénoncé dans plusieurs études[1574].



Une expérience similaire a été menée en Inde avec la startup israélienne **Cortica**[1575] qui exploite la vidéo-surveillance. Il en va de même pour l'expérience menée par **Palantir** à New Orleans sans prévenir la ville[1576] !

A contrario, une expérience d'usage de l'IA lancée en 2012 pour attribuer des niveaux de risques aux personnes arrêtées par la police de Chicago a été stoppée en 2020. En cause, les biais raciaux générés par des IA, du fait des données d'entraînement. Bref, le ProCog n'était pas la panacée[1577]. Il en va de même des diverses tentatives de remplacement des habituels détecteurs de mensonges par des variantes à base d'IA qui exploitent l'analyse d'expressions faciales, de la voix et autres paramètres biométriques[1578].

[1572] Voir The Growth Of AI Adoption In Law Enforcement par Kathleen Walch, juillet 2019. Et Clearview AI facial recognition customers reportedly include DOJ, FBI, ICE, Macy's par Corinne Reichert, mars 2020, Clearview AI is looking to expand globally, report says par Corinne Reichert, février 2020 et Clearview AI, la start-up qui pourrait mettre fin à votre vie privée par Amélie Charnay, janvier 2020.

[1573] Le digital evangelist Stéphane Mallard dans ses conférences, comme dans L'intelligence Artificielle - A l'aube de la disruption ultime, indique que ce système permet de prévoir les crimes à l'avance, avec la date et lieu. Ce n'est pas du tout la fonction de HART ! Comme de nombreux évangélistes du secteur, les exemples donnés qui relèvent d'une revue de presse de premier niveau sont souvent très exagérés dans leur portée et leur fonction réelle. En pratique, et pour ce qui concerne les anglais, cette fonctionnalité est anticipée pour 2030. Aujourd'hui et 2030, ce n'est pas la même chose ! Voir The real Minority Report: By 2030, police could use AI to predict and prevent crimes BEFORE they happen, septembre 2016. Les exagérations de ce genre sur l'IA sont très courantes.

[1574] Voir Pitfalls of Predictive Policing, septembre 2016, qui fait reference à l'étude de la Rand Corporation sur l'expérience menée à Chicago ici A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot.

[1575] Voir Crime-predicting A.I. isn't science fiction. It's about to roll out in India, avril 2018.

[1576] Voir Palantir has secretely been using New Orleans to test its predictive policing technology, février 2018.

[1577] Voir Chicago PD's predictive policing tool has been shut down after 8 years of catastrophically bad results par Cory Doctorow, janvier 2020.

[1578] Voir Lie detectors have always been suspect. AI has made the problem worse. par Jake Bittle, mars 2020.

Axon Enterprise, une filiale de **Taser International**, fournit des caméras de policiers qui aident ces derniers à faire du reporting automatique d'incidents et arrestations. Le système a été entraîné avec 30 Po de vidéos issues des caméras de 200 000 agents. En pratique, le système labellise les vidéos et floute les visages pour préserver leur anonymat si elles deviennent publiques par inadvertance. Mais tout cela ne fait pas l'unanimité, la reconnaissance de postures étant sujette à de nombreuses erreurs[1579].

UVeye (2016, Israël, $35,5M) utilise le traitement de l'image de manière originale, en scannant les dessous des véhicules pour détecter des menaces comme des bombes, armes ou caches diverses. Le système s'appuie sur une vision en 3D. Reste à récupérer de bonnes données d'entraînement (vidéo).

Evolv Technology (2013, USA, $83,8M) est un scanner corporel qui détecte les armes et explosifs avec moins de contraintes que les habituels scanners d'aéroports, à savoir que l'on n'est pas obligé de mettre de côté les objets métalliques. Le débit est de 600 personnes par heure ce qui reste encore faible (un toutes les six secondes). Le système utilise aussi la reconnaissance faciale pour détecter des personnes recherchées (vidéo).

L'**Université de Caroline du Sud** a même développé une solution qui prédit l'émergence de protestations violentes en surveillant les réseaux sociaux, un projet financé par le Ministère de la Défense[1580]. Dans un pareil cas, on s'approche des caractéristiques d'un état policier, réprimant toute forme de protestation comme c'est le cas en Russie. L'Internet est aussi utilisé comme un outil de répression et de régression des libertés dans les démocraties !

Deep Sentinel (USA, $23,4M) propose un système de caméra de surveillance domestique qui détecte les crimes avant qu'ils interviennent. Le système détecte toute activité anormale par rapport aux habitudes et déclenche l'alarme si besoin est. De là à en faire un PreCog de Minority Report, il ne faudrait pas exagérer !

Knightscope (2013, USA, $77,6M) développe une famille de robots de surveillance pour la police, à roulettes (le K5) ou sous la forme de 4x4 (le K7) qui est déployée aux USA mais dont les bénéfices ont l'air d'être bien survendus[1581]. Ces robots ne sont pas armés et n'ont pas de bras. Ils servent d'outil de surveillance d'appoint pour les policiers.

Ils peuvent aussi permettre d'alerter la police, pour peu qu'ils soient au bon endroit lorsque l'on en a besoin... et que son smartphone ne fonctionne plus ou a été volé.

La **Chine** expérimente un système équivalent qui ambitionne d'aller plus loin en tentant de prévoir où des crimes pourraient avoir lieu en suivant les déplacements de groupes de criminels connus[1582]. Un tel système ne peut fonctionner correctement que s'il dispose d'une base de donnée de ces suspects et s'il est capable de suivre leurs déplacements en temps réel. L'exploitation d'images de caméras de vidéo-surveillance peut servir à cela et à détecter des comportements suspects comme ceux des pickpockets[1583]. Des initiatives anglaises voisines ont aussi été lancées[1584] ainsi qu'au Japon[1585]. L'histoire n'indique jamais quelle est leur efficacité réelle et les effets de bord indésirables.

[1579] Voir The Trouble With Trusting AI to Interpret Police Body-Cam Video, par Dan Greene et Genevieve Patterson, novembre 2018.

[1580] Voir Social media posts may signal whether a protest will become violent, mai 2018.

[1581] Voir Knightscope: Peace Through Superior Technology, une vidéo d'octobre 2020 qui présente cette offre et fait miroiter une sécurité idyllique.

[1582] Voir China seeks glimpse of citizens' future with crime-predicting AI en juillet 2017.

[1583] Encore faut-il alors intervenir ! On n'a pas besoin d'IA pour identifier les pickpockets dans les lieux touristiques à Paris et ils ne sont pas arrêtés pour autant. Voir Chinese Police Arrest Suspect At Pop Concert Using Facial Recognition, mars 2018.

Dans le même domaine, citons cette expérience menée en Hongrie de détecteur de mensonge à base d'IA utilisé dans le contrôle des frontières pour détecter l'immigration illégale[1586]. Le système utilise un agent virtuel qui pose des questions. Le visage des répondants est analysé par l'IA avec des techniques probablement voisines de celles d'Affectiva. Les premiers résultats étaient exacts à 75%, ce qui est bien faible. L'équipe du projet **iBorderCtrl** pense que ce taux pourra atteindre 85%.

L'IA peut enfin aussi servir à résoudre des crimes non élucidés. Elle permet par exemple d'améliorer la détection d'une arme à feu utilisée dans un crime à partir de simples traces et résidus laissés par les tirs[1587].

Fiscalité

Le machine learning peut servir à détecter la fraude fiscale. Il s'appuie sur des méthodes voisines sur le principe des systèmes de détection de fraudes dans les banques et les assurances[1588]. Les outils utilisés à la Direction Générale des Finances Publiques en France exploitent diverses bases de données et le machine learning identifie des fraudeurs suspects qui sont ensuite contrôlés par des agents humains du fisc. Cela s'applique aussi bien aux particuliers qu'aux entreprises. On peut espérer que cela ne génère pas trop de faux positifs et le harcèlement de contrôle fiscal associé ! 640M€ auraient ainsi été récoltés par le fisc sur les trois premiers trimestres de 2019.

Les méthodes s'améliorant régulièrement avec l'exploitation de données diverses captées sur le terrain (comme récolter les plaques d'immatriculation de voitures de luxe dans les stations de sport d'hiver haut de gamme) et sur Internet (scan des photos de villégiature publiées dans les réseaux sociaux), elles titillent la société qui peut aller jusqu'à se rebiffer[1589]. En 2019, le Fisc français lançait avec Accenture une expérience dans trois départements de détection de piscines non déclarées via l'analyse automatique de photos satellites et de données cadastrales[1590].

Dépendant aussi de Bercy, les Douanes peuvent aussi faire largement appel au deep learning pour identifier diverses formes de fraudes aux importations. Pour analyser automatiquement les scans aux rayons X de colis suspects, identifier des contrefaçons via un scrapping des données de sites web et leur analyse, pour détecter des anomalies temporelles dans les produits en transit, etc.

Recherche

Est-ce que l'IA va bouleverser la recherche voire remplacer les chercheurs ? Pour l'instant, pas vraiment. Les outils à base d'IA étendent cependant la panoplie des outils logiciels déjà exploités par les chercheurs.

Ils permettent notamment de créer des simulations plus sophistiquées de modèles physiques ou biologiques, de mieux exploiter les bases bibliographiques[1591], de résumer des publications scientifiques[1592] et d'identifier des corrélations entre données disparates dans plein de domaines.

[1584] Voir Minority Report is real : we can now foresee crime par Lisa Korrigane, 2018 et Grande Bretagne : la police va utiliser une IA pour prédire les crimes par Pierrick Labbe, 2018.

[1585] Voir Au Japon, une IA arrête les voleurs avant même qu'ils ne passent à l'acte par Laura Boudoux, mars 2019. Liée à la startup Vaak qui exploite les images de caméras de surveillance pour détecter les comportements suspects via l'analyse du langage corporel.

[1586] Voir An AI Lie Detector Is Going to Start Questioning Travelers in the EU, de Melanie Ehrenkranz, octobre 2018. On plaint les personnes qui seront victims de faux positifs de ce genre de système et seront bloqués longtemps au contrôle des frontières.

[1587] Voir Artificial intelligence could help crack previously unsolvable murder cases, novembre 2018 et les details dans Quantitative profile–profile relationship (QPPR) modelling: a novel machine learning approach to predict and associate chemical characteristics of unspent ammunition from gunshot residue (GSR), 2019 (13 pages).

[1588] Voir De l'intelligence artificielle au Fisc pour détecter les fraudeurs par Pierre Labbé, février 2019.

[1589] Voir L'Assemblée nationale apporte de nouveaux garde-fous au "Big Brother" de Bercy par Pierre Benhamou, novembre 2019.

[1590] Voir Le fisc expérimente sur l'IA avec Accenture par Louis Adam, décembre 2019.

[1591] Voir With little training, machine-learning algorithms can uncover hidden scientific knowledge, juillet 2019 et Quand l'intelligence artificielle vient au secours de la découverte scientifique par Arnaud Moign, août 2019.

En **astronomie**, des réseaux de neurones entraînés sur des modèles d'observation servent à simuler la dynamique de l'Univers et son évolution en fonction des paramètres physiques qui le régissent[1593]. D'autres les utilisent pour mieux prévoir les risques de collision d'astéroïdes avec la Terre[1594], pour détecter les cratères qui se forment sur Mars[1595], pour classifier des supernovaes[1596], ou prédire les éruptions d'étoiles naines[1597].

En **climatologie**, en complément des habituelles techniques de simulation, l'IA peut aider à prédire la température au contact air-mer[1598], à détecter les plaques de glace en Arctique[1599], à prévoir des feux de forêts en Californie[1600] voir à développer des modèles évolutionnaires d'écosystèmes[1601]. L'IA est aussi mise à contribution pour faire des prévisions météo[1602].

En **laboratoire**, le deep learning peut servir à automatiser les expériences répétitives, notamment en biologie[1603].

Comme dans le reste des usages de l'IA, les exemples de son application dans la recherche relèvent toujours d'outillage ad-hoc réalisés au cas par car, besoin par besoin, science par science. C'est presque toujours du sur mesure.

Renseignement et défense

L'IA a évidemment de nombreuses applications dans le vaste secteur de la défense et du renseignement[1604]. Ce sont d'ailleurs de gros financeurs des recherches dans ce champ scientifique et technologique, notamment via l'agence DARPA du Pentagone, qui finance des concours, challenges et appels à projets de R&D appliquée.

Défense

Ceci active de nombreux laboratoires d'universités, startups et entreprises établies. En juin 2018, le Département de la Défense US lançait le **JAIC** (Joint Artificial Intelligence Center) pour piloter la recherche appliquée sur l'IA avec un budget de $1,7B étalé sur cinq ans.

[1592] Voir Can science writing be automated? par Malewar Amit, avril 2019.

[1593] Voir The first AI universe sim is fast and accurate—and its creators don't know how it works par Thomas Sumner, juin 2019 et D3M : l'intelligence artificielle qui simule à la perfection l'univers par Camille Zaghet, juin 2019.

[1594] Voir Des astéroïdes qui pourraient entrer en collision avec la Terre identifiés par un réseau neuronal par Nathalie Mayer, février 2020.

[1595] Voir AI is helping scientists discover fresh craters on Mars par laNASA, septembre 2020.

[1596] Voir Artificial intelligence classifies supernova explosions with unprecedented accuracy, décembre 2020.

[1597] Voir Astronomers enlist AI in the search for 'lethal' baby star eruptions par University of New South Wales, octobre 2020.

[1598] Voir Recurrent Neural Network Architecture Search for Geophysical Emulation par Romit Maulik et al, avril 2020 (13 pages).

[1599] Voir Artificial intelligence could revolutionize sea ice warnings par UiT The Arctic University of Norway, juin 2020.

[1600] Voir This startup can predict where the next California fire will start par Adele Peters, octobre 2020. À propos de la startup **Kettle** (2019, USA, $4,7M), un réassureur d'assurances qui aide les assurances à évaluer les risques liés au climat avec une solution de deep learning de type Swarm Neural Networks exploitant les données satellite et météo.

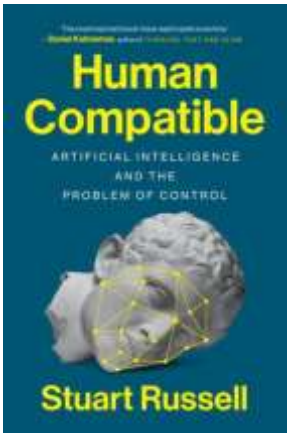
[1601] Voir Artificial intelligence helps scientists develop new general models in ecology par University of Helsinki, juin 2020. A partir d'une variante d'algorithmes génétiques, les régressions symboliques.

[1602] Voir AI model shows promise to generate faster, more accurate weather forecasts par Hannah Hickey, décembre 2020.

[1603] Voir Artificial-intelligence-driven scanning probe microscopy par A. Krull, mars 2020, puis Ce robot scientifique a réalisé près de 100 000 expériences en seulement un an par Mathilde Rochefort, 2019 et A robotic Intelligent Towing Tank for learning complex fluid-structure dynamics par D. Fan, novembre 2019 (13 pages).

[1604] Voir Artificial Intelligence and National Security de Greg Allen et Taniel Chan, 2017 (132 pages) et Artificial Intelligence and National Security de Daniel Hoadley et Nathan Lucas, 2018 (42 pages).

Ce JAIC dépend du DSI du DoD et non pas de la DARPA[1605]. Il serait bon que ces initiatives ne confient toutefois par l'arme nucléaire à une IA hors de contrôle de l'Homme[1606].



La France lançait un programme voisin en avril 2019 en allouant un budget de 100M€ de 2019 à 2025 pour déployer de l'IA au Ministère des Armées. Cela couvre différents domaines comme l'aide à la décision et à la planification, le renseignement, le « combat collaboratif », la robotique, la cyberdéfense, la logistique et la maintenance. 200 spécialistes en IA doivent être recrutés d'ici 2023[1607].

La robotique est déjà très largement utilisée, que ce soit avec des **robots de déminage** déjà opérationnels en Irak ou en Afghanistan (mais ils sont télécommandés) et surtout avec les **drones aériens**, eux-aussi télécommandés mais dotés d'outils de reconnaissance de leur environnement et de pilotage automatique. Ce sont soit des drones de reconnaissance, soit des drones lanceurs de missiles, éventuellement en mode essaim (swarm).

On arrive depuis déjà quelques années à faire atterrir des drones de manière autonome sur des porte-avions mais la part d'IA qui les anime n'est pas forcément énorme[1608]. En 2020, une IA était capable de piloter un avion de chasse plus efficacement qu'un pilote humain dans un simulateur.

Nous avons aussi vu dans la partie sur la marine que les navires de guerre autonome commencent à apparaître même s'il est probable que leur autonomie sera toutefois limitée.

Après encore quelques années de tests, l'US Army doit mettre en service son tank élecrique autonome **Ripsaw M5** produit par **Textron Systems** et **FLIR Systems**, ce dernier fournissant le système de vision nocturne ainsi qu'un système de conduite autonome exploitant quatre caméras à 360°. En pratique, il sera surtout utilisé en mode télécommandé. Et il n'est qu'hybride électrique, l'électricité provenant d'un moteur diesel embarqué !

Les robots tueurs entièrement autonomes ne sont pas encore opérationnels, mais rien n'empêchera les pays en roue libre d'en créer, malheureusement, même si comme le chercheur anglais Stuart Russell en fait la promotion, l'ONU déclare ces robots « illégaux »[1609].

[1605] Voir Establishment of the Join Artificial Intelligence Center, juin 2018 (2 pages).

[1606] Voir Pairing AI and nukes will lead to our autonomous Doomsday par Lori Esposito Murray, novembre 2018.

[1607] Voir La France dit non aux robots tueurs mais oui à l'IA militaire par Pierrick Labbe, 2019 et 100 millions d'Euros par an pour l'intelligence artificielle dans la défense, avril 2019.

[1608] Voir US lands drone on aircraft carrier at sea - video par ITN, 2013.

[1609] Voir Risques des robots tueurs, armes autonomes et les solutions pour y faire face de Dimitri Carbonnelle, mai 2018.

Il dirige le Center for Human Compatible AI (CHAI) de Berkeley et a publié l'ouvrage **Human Compatible** en octobre 2019 (349 pages)[1610]. Le Royaume-Uni n'hésite pas à envisager de d'équiper dès 2030 de soldats robots même s'il est probable qu'il s'agisse encore de chimères à cet horizon[1611].

En Corée du Sud, des robots armés de surveillance de la frontière avec la Corée du Nord ont été déployés, les SGR-A1 fabriqués par **Samsung** (ci-contre, sans le fusil-mitrailleur dont ils sont équipés).

De manière plus générique, l'IA doit pouvoir servir à consolider et organiser les informations qui alimentent les combattants, d'accompagner la prise de décision avec des outils d'analyse rapide de la situation notamment visuelle.

Renseignement

Dans le renseignement et l'analyse des interceptions de télécommunications, l'IA est évidemment utilisée, notamment avec des outils de traitement du langage. D'autres outils, basés sur du machine learning, permettent de détecter des bizarreries comportementales. Cela peut faire appel à du machine learning non supervisé servant à identifier des clusters atypiques de comportements. Les comportements peuvent concerner les déplacements ou les communications de personnes ou de groupes de personnes surveillées. Évidemment, ce n'est pas très bien documenté avec de belles études de cas de startups. La **NSA** comme la **DGSE** ou le **GCHQ** de nos voisins anglais ne sont pas du tout bavards sur la question.

L'IA peut ainsi servir à identifier des terroristes potentiels en fonction de profiling exploitant diverses sources de données, issues notamment de l'espionnage des communications électroniques. Ces outils utilisent du machine learning, de la PCA (Principal Components Analysis) pour identifier les paramètres permettant de les distinguer du reste de la population. L'IA sert à trouver des aiguilles dans de vastes bottes de foin.

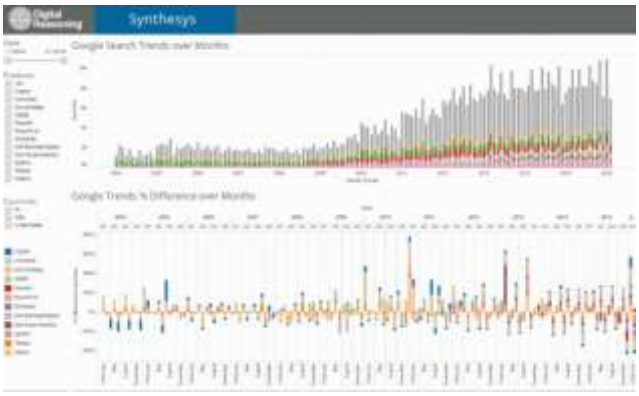
Palantir (2004, USA, $20B) est l'un des éditeurs de logiciels les plus connus du secteur. La société propose une suite logicielle destinée aux services de renseignement, de contre-espionage, anti-terroristes, des armées ainsi que de lutte contre diverses formes de fraudes dans le secteur de la finance. Sa suite logicielle Gotham d'analyse de données (data analytics) intègre des outils de visualisation de données sous forme de graphes, cartographies, explorateur d'objets, liés à des outils d'analyse de données à base de machine learning et de deep learning, le tout dans un environnement de travail collaboratif. C'est l'archétype de la solution intégrée faisant de la big data. Gotham cible les applications militaires et de renseignement. Son équivalent civil est Foundry. Il comprend aussi des outils logiciels d'acquisition de données exploitant les APIs ouvertes du marché. De son côté, Apollo délivre les services de Palantir en mode cloud. En fait, l'activité commerciale de Palantir est largement à dominante civile.

L'éditeur a des clients dans divers secteurs de l'industrie comme dans la santé (Merck, UK NHS pour la lutte contre la covid-19), les assurances (Zurich Insurrance), la construction aéronautique (Airbus) et même dans le suivi de flotte maritime. Palantir est utilisé hors des USA, ce qui génère de nombreuses questions de souveraineté. Mais ces craintes sont à relativiser tant que ces logiciels sont utilisés « on-premise » dans les data-centers des clients. Palantir faisait environ $1B de chiffre d'affaire en 2020 ce qui en fait un éditeur de logiciels de taille intermédiaire.

[1610] Il y décrit les manières de faire cohabiter les robots et les humains et en particulier comment interpréter et satisfaire les besoins de ces derniers en les plaçant dans leur contexte.

[1611] Voir 'Robot soldiers could make up quarter of British army by 2030s' par Dan Sabbagh, novembre 2020.

Digital Reasoning (2000, $73M) a été créée par des anciens d'Oracle et de la CIA (entre autres provenances) et est financée par In-Q-Tel, le fonds d'investissement de cette dernière. Sa solution d'analyse de données est utilisée par le renseignement et la défense US ainsi que dans la finance. Comme celle de Skymind, sa solution Synthesys est en Java et ouverte. Elle permet d'analyser des données structurées et non structurées, y compris des conversations téléphoniques.



Elle sert à détecter des comportements anormaux dans les communications électroniques et à prévoir les intentions de personnes surveillées. C'est donc probablement un outil exploité par la NSA dans la gestion de ses interceptions (PRISM & co).

Fifth Dimension (2014, Israël) est un concurrent de Palantir qui fournit des services d'analytics à base de machine learning et deep learning aux services de renseignement et d'enquêtes policières.

Primer.ai (2015, UK, $58M) propose des outils multilingues permettant de détecter des campagnes de désinformation touchant des pays ou des entreprises. Ils sont exploités par le Pentagone aux USA, par exemple pour identifier des opérations de déstabilisation provenant par exemple de la Russie. Le système analyse les sources d'information publiques - ou pas - sur des thèmes prédéfinis, principalement textuelles[1612]. Ici, l'IA est donc mise à contribution pour détecter des fake-news d'origine humaine. La startup a aussi été financée par In-Q-Tel.

Enfin, les GANs peuvent servir à créer de **faux personnages**, un peu dans la lignée des « légendes » de la série *Le Bureau des Légendes* de Canal+. C'est ce qui a été découvert sur un faux profil LinkedIn généré avec un GAN de création de visages composites. Il était associé à un faux compte d'une certaine Katie Jones, reliée à une cinquantaine de personnes du gouvernement US[1613]. Et encore, la technique pourrait être perfectionnée en créant la partie historique et texte de ces personnages également à partir de GAN, mais textuels.

L'IA est aussi exploitée dans la prévision d'événements et de risques géopolitiques comme chez **Predata** (2015, USA, $3,25M).

C'est un bon exemple d'usage sophistiqué de l'IA pour un exécutif, et qui pourrait aussi s'appliquer aux prévisions dans la création de politiques économiques. Mais celles-ci sont toujours créées en fonction d'un croisement entre idéologies partisanes, méthode Coué et en jouant avec les contraintes de l'administration.

[1612] Voir The AI Company Helping the Pentagon Assess Disinfo Campaigns par Will Knight, octobre 2020.

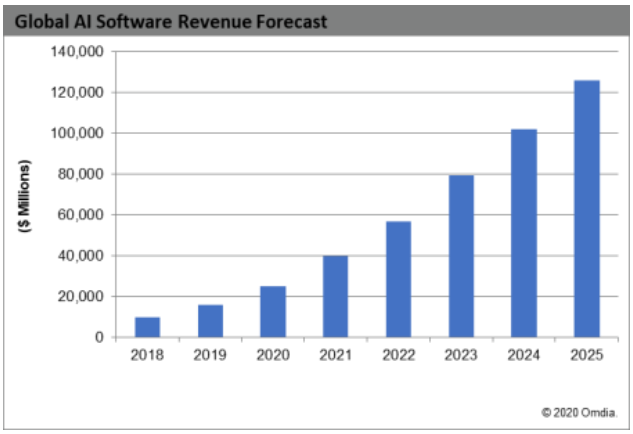
[1613] Voir The LinkedIn user who wasn't there: AI-generated 'person' was used to spy par Steven Musil, 2019.

Acteurs de l'IA

Une fois que l'on a fait le tour des algorithmes et technologies de l'IA et de ses usages potentiels souvent issus de nombreux laboratoires de recherche, il faut passer par la case « action » et avancer. Pour cela, les entreprises ont besoin de faire appel à un panaché d'acteurs qu'il nous faut examiner : les grandes entreprises du numérique dont les GAFAMI, les startups et les entreprises de services du numérique (ESN). Nous allons ici examiner ces catégories d'acteurs, leurs offres, forces et faiblesses. Nous terminerons avec un tour de l'écosystème français de l'IA, en y intégrant la recherche[1614].

Ces dernières années, les prévisions de croissance du marché de l'IA sont allées bon train. Au départ, elles correspondaient à une véritable nouvelle catégorie bien à part de logiciels. Mais, progressivement, elles perdent tout sens car l'IA est intégrée dans presque toutes les catégories de logiciels métiers et d'infrastructure. Dans son état actuel, l'IA n'est pas véritablement une catégorie de produits. C'est un patchwork de techniques plus ou moins nouvelles intégrées dans un nombre grandissant de solutions logicielles[1615].

Selon une prévision de janvier 2020, le marché du logiciel d'IA devrait passer de $10,1B en 2018 à $126B en 2025 (source : Omdia). En pratique, ce n'est pas une taille de marché incrémentale, mais plutôt la part du marché global du logiciel qui va intégrer de l'IA, celui-ci étant situé autour de $487B en 2020 (source Statista). Les marchés liés à la vision et au traitement du langage sont assez faciles à détourer. Par contre ceux du machine learning autour de données structurées sont enfouis dans les marchés des logiciels métiers. Une autre prévision de début 2020 mettait en tout cas en premier ces applications exploitant du machine learning dans ses prévisions (*ci-dessous* à droite)[1616].



Grandes entreprises du numérique

Les grands acteurs occidentaux du numérique sont tous très impliqués dans l'IA pour améliorer leurs solutions. Nous avons en tête les GAFAMI : Google, Amazon, Facebook, Apple, Microsoft et IBM. Microsoft et IBM sont intégrés dans ces GAFA étendus du fait de leur double forte présence dans le système d'information des entreprises et par leur panoplie généraliste de développement d'applications d'IA. Nous y ajoutons ici quelques autres grands éditeurs de logiciels d'entreprises comme Oracle, SAP et Salesforce.

[1614] Voir Digital Index, AI Index de Stanford, 2019 (291 pages) qui compile un très grand nombre d'indicateurs mondiaux sur l'IA au niveau de l'activité scientifique comme économique.

[1615] Voir Why "AI" Is A Fraud par Stephen McBride, septembre 2020, qui décrit bien cette diversité.

[1616] Voir Roundup Of Machine Learning Forecasts And Market Estimates, 2020 par Louis Columbus, janvier 2020.

Tous les GAFAMI ne jouent pas le même rôle dans les grandes entreprises. Seuls IBM, Microsoft, Amazon et dans une seconde mesure Facebook et Google, proposent des plateformes et solutions adaptées aux besoins des grandes entreprises.

IBM et Microsoft sont les entreprises investies dans l'IA depuis le plus longtemps. L'un des initiateurs du Summer Camp de Darmouth en 1955 était Nathaniel Rochester, un chercheur d'IBM. Et Microsoft a créé son laboratoire de recherche en 1991, principalement dédié aux avancées de l'IA, traversant des époques où cela n'était pas du tout à la mode, notamment dans le domaine complexe du traitement du langage.

Les GAFAMI publient en open source presque tous leurs outils logiciels d'IA. Ce sont des commodités. Il n'est pas évident de départager les outils de reconnaissance d'image et du langage d'IBM, Google, Amazon et Microsoft. Les différences se situent plus dans leur manière d'aborder le marché des entreprises. On peut aussi les situer dans la capacité de ces acteurs à fournir des outils de création de solutions d'IA accessibles au plus grand nombre avec un bon modèle économique. Enfin, peut entrer en ligne de compte la capacité à entraîner des modèles de deep learning génériques et géants comportant des centaines de milliards de paramètres destinés surtout au traitement du langage et de la vision artificielle.

La réussite dans l'IA comprend quelques ingrédients de base : des plateformes logicielles intégrées et ouvertes, des ressources en cloud éventuellement enrichies par des architectures matérielles propriétaires différenciées, l'accès à des données d'entraînement de qualité, des partenaires adoptant la plateforme, ou la capacité interne à gérer des projets clients et enfin, surtout, des talents, qui sont de plus en plus difficiles à dénicher et/ou à former.

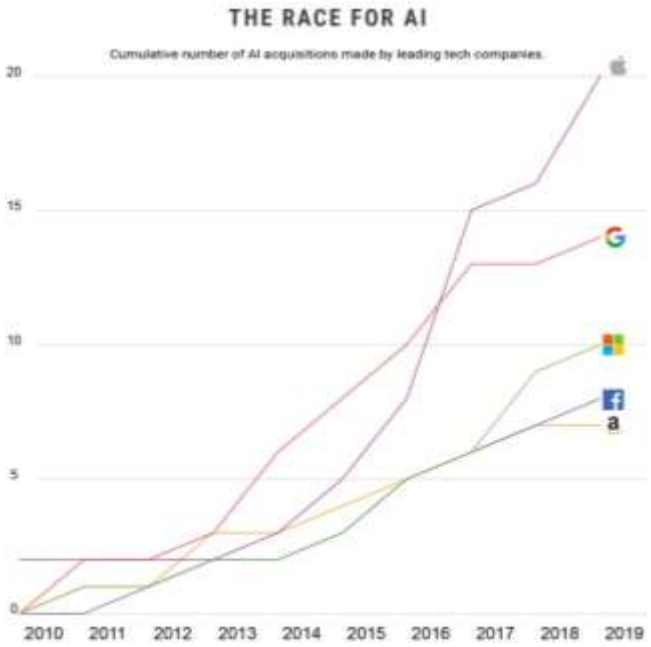
Sur les grands marchés horizontaux, ces GAFAMI et leurs équivalents chinois sont déjà les leaders mondiaux de l'IA[1617]. Il reste cependant encore des places à prendre pour des acteurs positionnés sur des marchés verticaux que ces grands acteurs ciblent mal. Et d'ailleurs, plus un marché est fragmenté, moins les GAFAMI risquent de s'y attaquer. C'en est au point où de nombreux usages de l'IA sont complètement délaissés par les grands acteurs, faute d'économies d'échelle.

J'évoquerai le cas des grands acteurs du numérique chinois à part dans la rubrique sur la géopolitique de l'IA. Ceux des acteurs à observer de plus près au sujet de l'IA sont **Huawei** et **Baidu**. Le premier avec sa stratégie d'IA de bout en bout allant des infrastructures d'opérateurs télécoms jusqu'aux chipsets de smartphones (Kirin 970/980/990) et même pour serveurs, et le second avec sa stratégie de plateforme d'agent vocal DuerOS et surtout Apollo pour le pilotage de véhicules autonomes.

L'autre benchmark habituel des grands groupes *pure players* du numérique consiste à observer leurs acquisitions. Comme il y a une guerre de talents dans l'IA, c'est un indicateur de leur montée en puissance.

[1617] Voir Why AI consolidation will create the worst monopoly in US history, août 2016.

De ce côté-là, **Google** était en tête de peloton jusqu'en 2017 mais **Apple** a depuis pris le dessus comme plus gros acquéreur de startups[1618]. Ces acquisitions sont soit technologiques, soit de l'ordre de l'*acquihire*, un barbarisme décrivant le recrutement de talents au prix fort. Il faut aussi intégrer quelques investissements « préventifs ».



Les GAFAMI adoptent aussi une double approche de **plateforme** pour attirer des développeurs d'applications avec des frameworks le plus souvent open source et générer des économies d'échelle et d'**intégration verticale** pour capter une partie aussi grande que possible de la valeur ajoutée, qu'il s'agisse de sa dimension technique ou des usages, parfois jusqu'à cibler des marchés verticaux.

Cette intégration verticale couvre de nombreux domaines comme les chipsets dédiés à l'IA sur serveurs ou embarquée (IBM, Google, Microsoft, Intel, Apple), les applications et données grand public (tous sauf IBM), les données métier (surtout IBM), la mobilité (Google, Apple, Facebook), la réalité virtuelle (Google, Facebook, Microsoft), la création et/ou la distribution de contenus (Google, Amazon, Apple, Facebook) et plus rarement, les services, le conseil et l'intégration (surtout chez IBM et Oracle).

Enfin, ces entreprises sont certes américaines, mais elles ont toujours une forte ambition internationale. Propulsée économiquement par leur marché intérieur US, elles s'installent ensuite rapidement dans les pays Européens pour s'y imposer. Cela génère parfois des résistances. Il faut donc bien interpréter les tactiques des GAFAMI lorsqu'ils font leur marketing dans les pays comme la France, y installent des centres de recherche et mènent diverses actions de lobbying servant de tranquillisant pour les élites politiques, scientifiques et économiques.

Ils investissent localement dans la recherche (Microsoft, Google, Facebook, mais aussi Fujitsu et Naver), dans l'éducation (idem, Cisco…) et nouent des partenariats avec les grandes entreprises qui se jettent dans la gueule du loup un peu trop facilement alors qu'il ne s'agit souvent que de relations clients/fournisseurs un peu galvaudées et très intéressées[1619]. La souveraineté technologique se perd petite bouchée par petite bouchée !

Voici un petit résumé comparatif de la situation des GAFAMI et autres grands acteurs du numérique :

- **IBM** est un acteur très présent dans les grandes entreprises avec sa plateforme logicielle « couteau suisse » IBM Watson et son activité de services associés, sans compter ses offres verticales comme dans la finance. Et la société est investie dans l'IA depuis ses débuts, en 1955 avec Nathaniel Rochester. Malgré cela, IBM a du mal à rentabiliser son investissement dans Watson. Son CA est en baisse et la concurrence est rude dans les services autour de l'IA. Qui plus est, la traction de Watson n'est pas très bonne chez les développeurs qui ont adopté en priorité les outils de développement de Google (TensorFlow), Facebook (Caffe, PyTorch) si ce n'est de Microsoft. Et pour couronner le tout, IBM va se scinder en deux en 2021 !

[1618] Voir The Race For AI: Here Are The Tech Giants Rushing To Snap Up Artificial Intelligence Startups, septembre 2019.

[1619] Voir Atos et Total s'allient à Google dans l'intelligence artificielle, avril 2018 et Carrefour s'allie avec Google sur l'e-commerce, l'innovation et la bureautique, juin 2018.

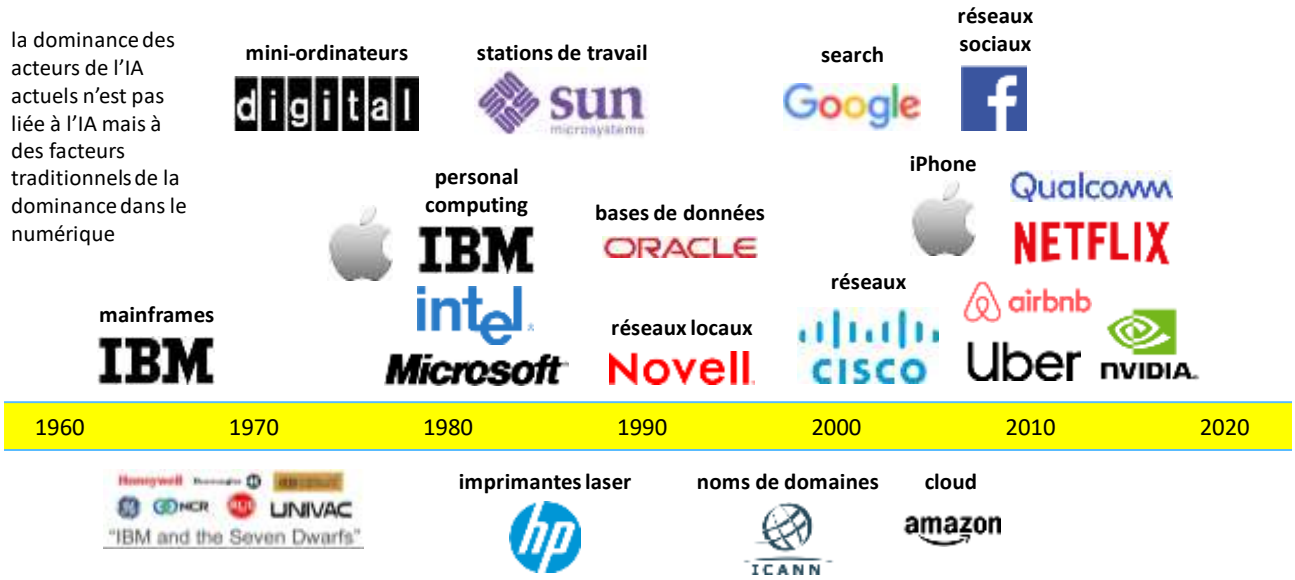
- **Google** est un acteur dominant du numérique grand public couvrant de larges pans de la vie des utilisateurs (recherche, email, mobile, TV, maison connectée). Cela lui permet de capter d'énormes volumes de données pour alimenter « ses IA ». Il domine aussi le développement de solutions d'IA avec son framework open source TensorFlow. C'est un opérateur de cloud qui a l'ambition de servir aussi les entreprises, concurrençant de ce point de vue Amazon et Microsoft. Il développe ses propres processeurs pour le deep learning aussi bien côté cloud que pour l'embarqué, les TPU.
- **Amazon** est le leader de la distribution en ligne et du cloud d'entreprise. Il s'est taillé une bonne place dans les assistants vocaux avec Alexa, qui est concurrencée par Google Assistant. Sa plateforme logicielle d'IA semble toutefois dans son ensemble moins influente que celle de Google. C'est surtout le leader mondial des ressources en cloud pour les entreprises. Et il met maintenant en production ses propres processeurs pour le deep learning.
- **Microsoft** propose une plateforme logicielle complète d'IA couvrant le machine learning et le deep learning, sur les grands pôles que sont le traitement des données, du langage et de l'image, ainsi qu'une excellente activité de recherche fondamentale. C'est aussi un important opérateur de cloud, derrière Amazon. L'éditeur souffre cependant d'un déficit marketing dans le domaine, notamment auprès des startups qui sont peu nombreuses à adopter ses SDK et APIs.
- **Facebook** domine les réseaux sociaux et la communication mobile, des logiciels qui exploitent de plus en plus d'IA, aussi bien dans le traitement du langage (Messenger) que de l'image (WhatsApp et Instagram). L'influence des développeurs passe par ses APIs pour créer des chatbots pour Facebook Messenger et par le framework PyTorch qui monte en puissance, au moins chez les chercheurs en IA.
- **Apple** est une société très intégrée verticalement dont l'approche plateforme, surtout en cloud, est moins prégnante dans l'industrie. iOS étant un passage obligé, il est mécaniquement présent chez de nombreux développeurs. Apple ne joue pas un grand rôle dans l'IA des entreprises. Il a intégré le support du deep learning dans ses processeurs depuis 2017. Après l'iPhone et l'iPad, cela touche même le Macintosh, depuis l'abandon d'Intel et la sortie du processeur M1 en 2020.
- **Intel** et **Nvidia** sont les deux leaders des processeurs pour serveurs et dans une certaine mesure dans l'embarqué, même si ce dernier marché est très fragmenté avec de nombreux acteurs. Intel a une offre comprenant une myriade de CPU et processeurs neuromorphiques pour serveurs et l'embarqué, certains étant adaptés au traitement du langage et d'autres, de l'image. Nvidia a une offre assez cohérente de GPU avec des chipsets intégrant des unités de traitement classiques (ALU) et des tenseurs (multiplicateurs de matrices) qui sont adaptés au machine learning et au deep learning. Ils dominent ce marché de loin, notamment dans l'équipement du cloud et des supercalculateurs.
- **Oracle**, **SAP** et **Salesforce** ont mis l'IA à leur menu mais font moins parler d'eux du fait de leur positionnement entreprise. Salesforce est probablement le plus avancé des trois dans les usages de l'IA même s'il peut avoir tendance à survendre Einstein et à maintenir quelque flou sur les données qui servent à l'entraîner.

Toutes choses que nous allons examiner un peu plus en détail dans ce qui suit, acteur par acteur !

La question se pose de savoir si leur dominance de leurs marchés respectifs s'affirme grâce à l'IA. C'est probablement le cas mais ça l'est de nombreuses autres technologies qu'ils ont intégrées progressivement, notamment autour des outils du cloud et de la mobilité[1620]. La dominance est accentuée par la taille des modèles de deep learning qu'ils peuvent entraîner pour améliorer la qualité de

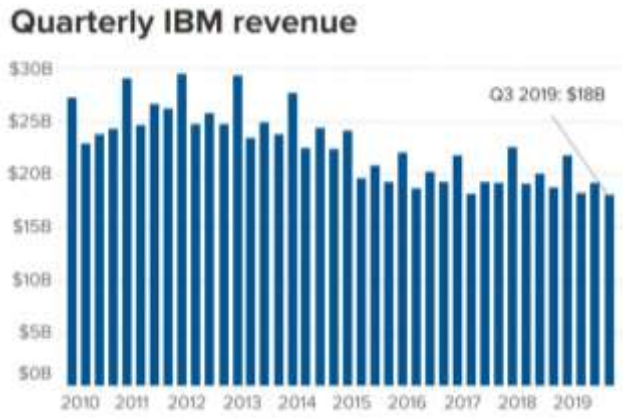
[1620] Voir Emerging AI Will Drive The Next Wave Of Big Tech Monopolies par Ed Stacey, octobre 2020, qui fait notamment référence à un rapport de la Chambre des Représentants US faisant un état des lieux de la situation monopolistique atteinte par chacun des GAFA : Investigation of competition in digital markets, Majority staff report and recommendations, Subcommittee on antitrust, commercial and administrative law of the committe on the judiciary, octobre 2020 (455 pages).

leurs services. Cela rentre dans le cadre plus général des notions d'économies d'échelle et d'effets de réseaux qui favorisent les leaders du marché dans tout un tas de domaines et ceux qui sont issus des USA, ce depuis des décennies comme l'illustre la timeline suivante remontant aux années 1960.

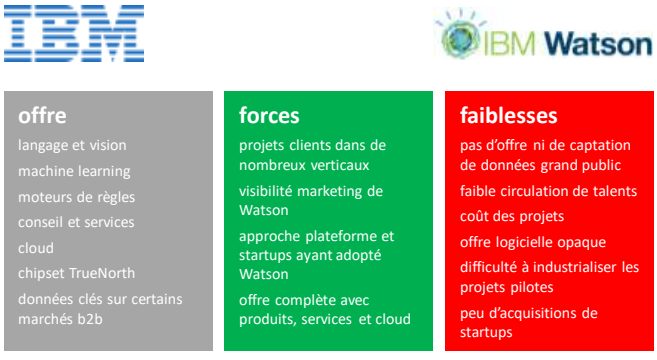


IBM

IBM est l'un des premiers grands acteurs du numérique qui ait mis le paquet sur l'IA relativement tôt, au moins à partir de 1996. Il a ensuite tout misé, d'un point de vue marketing, sur la marque IBM Watson qui regroupe les principales briques logicielles d'IA. IBM articule l'IA autour du concept d'informatique cognitive. Comme le montre cette évolution du chiffre d'affaires d'IBM, cela ne leur a pas permis de faire de la croissance. Ils se sont délestés de plusieurs activités logicielles spécialisées et l'activité est « flat » depuis 2016[1621].



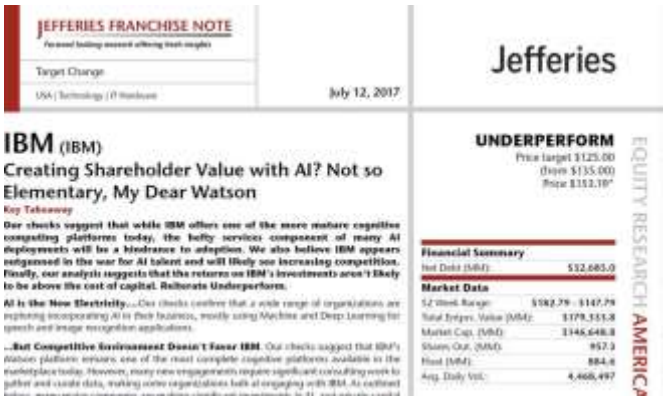
Une analyse au vitriol de 53 pages de juillet 2017 par la banque d'investissement Jefferies[1622] décrivait bien les enjeux d'IBM vus par la lorgnette des actionnaires : Watson est un magnifique artifice de communication, mais sa traduction en avantage compétitif n'est pas évidente pour l'entreprise dont la stratégie est tiraillée entre celle de prestataire de services et d'éditeur de logiciel. Et elle fait face à des concurrences multiformes.



[1621] Voir IBM stock falls on revenue miss par Jordan Novet, CNBC, octobre 2019.

[1622] Voir IBM Creating Shareholder Value with AI? Not so Elementary, My Dear Watson, Jefferies Franchise Note, juillet 2017.

Elle comprend les grandes sociétés de services et intégrateurs (CapGemini, Atos, Orange Business Services pour la France), les grands éditeurs de plateformes (Google, Microsoft, Oracle…) et les fournisseurs de cloud (Amazon et Microsoft). IBM a aussi bien du mal à attirer les talents qui sont aspirés par les GAFA et les startups de la Silicon Valley. Au point que début septembre 2017, IBM annonçait le lancement d'un laboratoire conjoint de recherche avec le MIT financé à hauteur de $24M par an sur 10 ans.


JEFFERIES FRANCHISE NOTE

Target Change

July 12, 2017

Jefferies

IBM (IBM)

Creating Shareholder Value with AI? Not so Elementary, My Dear Watson

UNDERPERFORM

Price target $125.00

(from $135.00)

Price $153.18*

Financial Summary

Net Debt ($MM): $52,685.0

Market Data

52 Week Range: $582.79 - $147.79

Total Enterp. Value ($MM): $179,533.8

Market Cap. ($MM): $146,648.8

Shares Out. ($MM): 957.3

Float ($MM): 884.4

Avg. Daily Vol.: 4,460,497

EQUITY RESEARCH AMERICA

En octobre 2020, IBM annonçait son projet de se scinder d'ici fin 2021 en deux entreprises, poursuivant des décennies de restructuration démarrées au début des années 1990 et qui se poursuivent encore aujourd'hui, avec un plan de licenciement de 10 000 personnes en Europe ayant fuité en novembre 2019 et affectant notamment l'Allemagne, le Royaume-Uni et la France (pour plus de 1200 postes). Nous aurons d'un côté un IBM conservant les offres logicielles, l'IA et les solutions de cloud autour de Red Hat faisant environ $59B de CA, et de l'autre, une nouvelle société pas encore nommée qui gérera l'activité de services et de conseil et totalisant environ $19B de CA. Bref, IBM va conserver l'activité générant des revenus récurrents et les meilleures économies d'échelle. Cela fera passer ces revenus récurrents à plus de la moitié du CA alors que jusqu'à présent l'activité des services représentait plus de la moitié du CA de l'IBM avant scission. Cela permettra au passage à IBM de développer plus sereinement des partenariats avec les nombreux acteurs des services informatiques dans le monde, en particulier, autour de leur riche boîte à outils Watson.

L'histoire

Dans les années 1960, IBM aurait stoppé brutalement ses travaux de recherche en IA par peur que les postes de managers soient remplacés par des machines. C'était aussi le résultat d'une remontée des clients dans les DSI qui avaient aussi peur de perdre leur poste. L'histoire a tendance à se répéter pour ce qui est des réactions du marché. On sous-estime d'ailleurs les effets de boucle rétroactive entre les promesses effrayantes que certains relaient sur l'IA et la réaction de rejet ou de prudence que cela peut générer dans les entreprises.

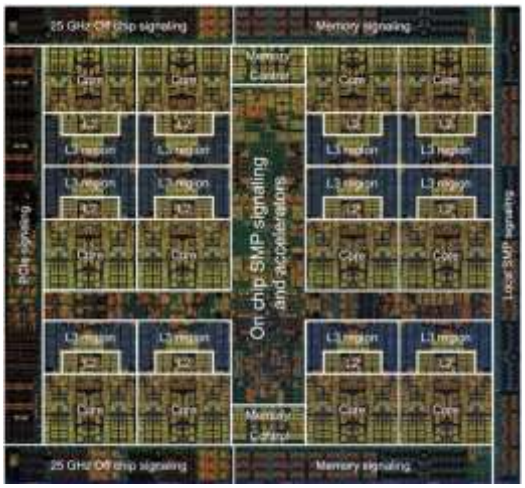
IBM a depuis fait sa mue de constructeur vers le métier d'éditeur de logiciels couplé à celui de prestataire de services à partir de 1993. IBM génère maintenant l'essentiel de son profit à parts égales entre logiciels et services. La synergie entre les deux métiers est plutôt bonne même si la branche services d'IBM travaille aussi avec les technologies concurrentes.

Ils savent déployer des solutions qui intègrent des logiciels d'Oracle, Microsoft, SAP, bref de tout, en fonction des contraintes du client.

La question reste cependant pour tout acteur du marché de ne pas rater les vagues technologiques. IBM ne s'en était pas trop mal sorti en 2000 en se positionnant dans le e-business. Sa campagne de communication martelait le rôle de fournisseur "one-stop-shop" pour ses clients.

IBM a petit à petit délaissé ses activités matérielles dans les machines de commodité. Le délestage s'est fait par étapes : les imprimantes avec la création de Lexmark en 1991, les PC cédés en 2004 au chinois Lenovo, et puis les serveurs PC cédés également à Lenovo, en 2014. Par contre, ils ont toujours misé sur les grandes architectures, dans la lignée de leur ligne historique de mainframes. D'où l'importance pour eux du HPC (High Performance Computing) et de l'intelligence artificielle. Il leur reste les mainframes de la série Z et les serveurs Power9 équipés du chipset du même nom (*ci-dessous*), avec 12 cœurs et produit en 14 nm FinFET sur wafers SOI (silicon on insulator) d'origine SOITEC.

Ils investissent aussi dans des chipsets neuromorphiques et dans l'informatique quantique. La recherche fondamentale d'IBM reste l'une des plus actives dans le secteur privé.



La première incartade d'IBM dans l'IA s'est manifestée au grand jour avec la victoire de l'ordinateur **IBM Deep Blue** (appelé initialement Deeper Blue) contre Gary Kasparov en 1997. Cela a contribué à relancer les recherches d'IBM sur l'IA dans les années 2000. La seconde grande étape a été la victoire d'IBM Watson au jeu **Jeopardy** en 2011. Jeopardy est une sorte de "Questions pour un Champion" inversé américain, sans Julien Lepers. Watson n'est pas infaillible.

Cette victoire fut un peu enjolivée et construite par la communication d'IBM qui au passage, a été pilotée à l'échelle mondiale par l'agence **Ogilvy**[1623]. Watson était au départ un projet de recherche baptisé BlueJay (2007) focalisé sur l'exploitation de gros volumes de données non structurées. Il s'intégrait dans la volonté d'IBM Research de s'attaquer à un grand défi, la réussite du fameux test de Turing. Watson était d'abord présenté comme un ordinateur, s'appuyant sur une architecture massivement parallèle à base 750 serveurs utilisant des processeurs Power7 octocœurs tournant à 3,5 GHz et totalisant 16 To de RAM.

IBM Watson est devenu une plate-forme logicielle, respectant en cela les canons de la réussite dans le numérique[1624]. Elle est proposée aux développeurs sous forme d'APIs (interfaces de programmation) dans le cloud. Watson s'appuie principalement sur la solution **DeepQA** d'IBM et le framework **Apache UIMA** (Unstructured Information Management Architecture) qui permet d'exploiter des données non structurées et de créer des chatbots. Suivent d'autres briques de traitement du langage et pour la vision artificielle.

En 2011, Watson devenait le sujet phare de la communication d'IBM autour de la notion de « cognitive computing », une appellation qui leur a permis d'éviter aux débuts celle de l'IA qui n'avait pas encore bonne presse. IBM englobe dans cette appellation presque toutes les applications qui gèrent… de l'information, qu'elles s'appuient ou pas sur des briques logicielles d'IA, y compris la gestion de bases de données transactionnelles. Depuis 2016, IBM utilise les deux. IBM organise aussi des conférences autour de Watson, d'abord "World of Watson" puis « Think »[1625].

En France, IBM se focalise notamment sur le développement des compétences autour de l'IA. Ils sont pour ce faire partenaires d'un bon nombre d'établissements d'enseignement supérieur.

[1623] Dans une partie intéressante et moins médiatisée (vidéo) organisée avec Miles O'Brien et David Gondek, l'un des créateurs de Watson, Watson ne sait pas indiquer pendant quelle décennie Klaus Barbie a été condamné ni indiquer sur quelle place de Dallas (Dealey Plaza) JFK a été assassiné, ni ce qu'est la vermiphobia (la phobie des vers) ou la ailuraphobia (phobie des chats), toutes ces informations étant pourtant disponibles sur Wikipedia. Il ne savait pas non plus identifier des recettes de cuisine en fonction de leurs composantes. Watson avait aussi du mal à répondre à des questions formulées avec peu de mots et comprenant des ambiguïtés ou des doubles sens. Tout est question de base de connaissances. Celle-ci comprenait 200 millions de pages de données structurées et non structurées représentant un total de 4 To, toutes chargées en mémoire pour assurer un temps de réponse rapide. C'était d'ailleurs bien injuste, car les joueurs humains n'avaient pas le droit d'accéder à Wikipedia sur leur mobile. À l'époque, les smartphones étaient déjà de la partie ! La mémoire verbale humaine est à peine de quelques Go. Elle ne peut pas concurrencer une mémoire de 4 To !

[1624] L'histoire est bien racontée dans IBM Watson: The inside story of how the Jeopardy-winning supercomputer was born, and what it wants to do next de Jo Best, 2015.

[1625] La dernière conférence IBM Think en présentiel avait lieu février 2019 à San Francisco. L'édition 2020 avait lieu en ligne en mai.

Ils comprennent l'**emlyon business school** avec qui ils développent un « GPS des compétences », une plateforme d'anticipation des tendances en matière d'emplois, de modèles de formations et d'accompagnement, l'**Ecole Nationale Supérieure de Cognitique** de Bordeaux, pour créer une chaire en Sciences Cognitives avec une formation en IA, l'**ESIEE Paris**, pour un autre chaire « IA et prescriptive analytics » et avec **Grenoble INP-ENSIMAG**, pour l'accès à des ressources de calcul en cloud. Tout cela, bien entendu, autour d'IBM Watson. En 2019, IBM lançait un partenariat avec l'**ESIEE** en créant une chaire d'entreprise en IA. Elle encouragera aussi les élèves à développer des logiciels quantiques avec la plateforme en ligne Q d'IBM. Enfin, ils ont ouvert un laboratoire à **Montpellier** sur l'IA et l'informatique quantique et un autre sur **Paris-Saclay**. Dans les deux cas, cela semble surtout être un centre technique pour mener des projets pilotes avec des clients et partenaires.

Les logiciels

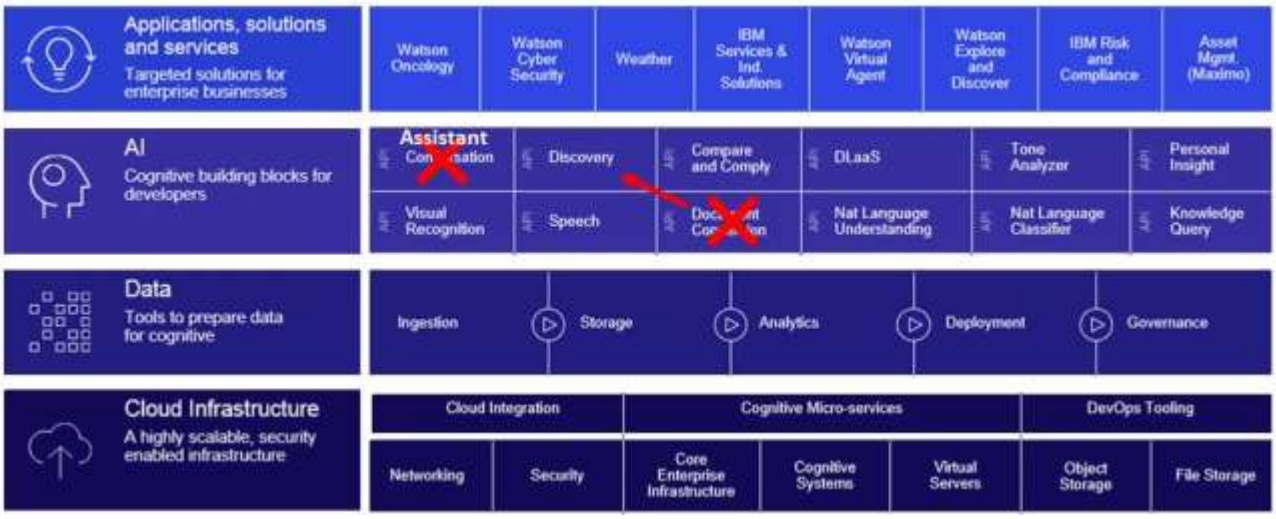
À chaque solution d'IA son assemblage de composants hétéroclites réalisé sur mesure pour répondre à un besoin. C'est particulièrement vrai d'IBM Watson. Ce dernier est un très bon coup business et marketing d'IBM, qui a réussi à simplifier partiellement un sujet très complexe.

Ils ont ainsi vulgarisé les capacités de Watson et pu cacher sa complexité, voisine de celle de l'architecture de WebSphere. IBM Watson est comme le fakir du célèbre sketch de Pierre Dac et Francis Blanche[1626] : dès que l'IA peut jouer un rôle dans un projet, « *il peut le faire* ».

Mais IBM Watson n'est ni un produit ni une solution ni « une IA ». C'est un attirail fait de nombreuses briques logicielles qu'il faut assembler et au-dessus desquelles il faut développer des solutions logicielles sur mesure pour en tirer quoi que ce soit. IBM passe d'ailleurs son temps à changer les noms de ses briques d'une année à l'autre !

Je n'ai pas réussi à trouver une version publique plus récente du schéma *ci-dessous* alors que Conversation devenait Assistant, que Document Conversion était intégré dans Discovery, sans compter le fait que le module Machine Learning ne figure pas dans ce schéma ni leur moteur de règles issu du rachat d'ILOG il y a 10 ans. Les modules « cognitifs » sont à l'exception de celui de la vision tous liés au traitement du langage.

[1626] Voir Le Sar Rabindranath Duval de 1956 à 7 minutes 44s qui comprend d'ailleurs le sketch « Le biglotron » qui constitue une excellente description prémonitoire d'IBM Watson dans à 9 minutes et 50 secondes qui date de 1958. Et une variante avec La voyante Arnica qui date de 1957, à partir de la cinquième minute.



IBM Watson est proposé aux développeurs de solutions sous la forme d'APIs REST[1627] qui permettent d'accéder à une large panoplie de services, qui sont intégrées dans la plateforme en cloud Bluemix.

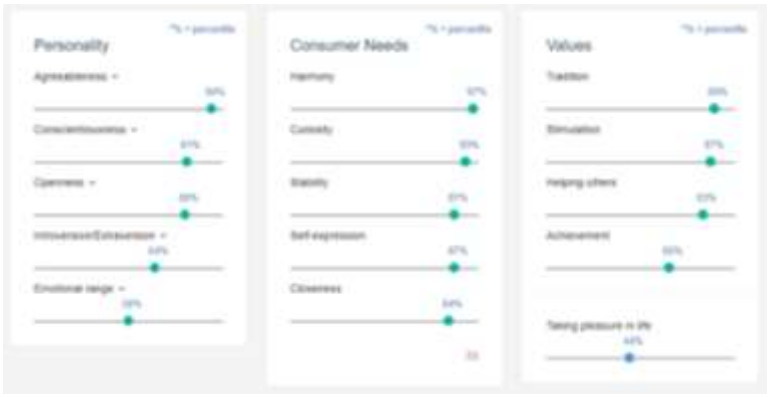
Commençons avec le gros morceau de ces APIs qui est lié au traitement du langage :

- **Assistant** (anciennement Dialog puis Conversation) permet de gérer des conversations scriptées pour des agents conversationnels, avec des arbres de décision. Ce genre d'outil est mis en œuvre dans les systèmes de chat des sites de commerce en ligne. Les dialogues générés sont généralement limités car préprogrammés[1628]. Le système peut être débrayé pour envoyer la conversation vers un agent humain lorsque le chatbot est perdu. Les références d'IBM dans les chatbots réalisés avec Assistant sont nombreuses avec notamment celui destiné aux agents du Crédit Mutuel de France, le Easy Button de Staples et Ask Mercedes.
- **Discovery** (anciennement Retrieve and Rank + Document Conversion) sert à ingérer et convertir tout document textuel (PDF, Word, HTML) pour les exploiter avec les services de Watson. Cela sert à l'alimentation de bases de connaissances. Le module Smart Document Understanding sert à interpréter la structure des documents.
- **Natural Language Understanding** sert à l'extraction de données et de connaissances dans des textes. Il permet aussi de classifier des documents et des pages web.
- **Natural Language Classifier** permet de classifier automatiquement des données textuelles, issues en général de questions posées par des clients en langage naturel. Cela les associe aux questions d'une base existante de questions. Cela permet aussi de classifier des mails pour identifier des spams. C'est un outil générique exploitable pour créer toutes sortes de solutions à base de traitement du langage. L'outil est exploitable dans le Knowledge Studio qui permet d'identifier des entités nommées dans des textes et de gérer leurs relations.

[1627] Les requêtes REST permettent à une application html d'interroger un serveur web. Ce sont des requêtes http comprenant des GET et des POST et renvoyant le résultat.

[1628] Voir ce tutoriel de développement de chatbot datant de début 2017, exploitant Conversation et d'autres briques logicielles d'IBM Watson : https://www.ibm.com/developerworks/library/cc-cognitive-chatbot-watson/index.html. Ainsi que Integrate Watson Assistant With Just About Anything par Mitchell Mason, avril 2018.

- **Personality Insights** analyse la personnalité d'un utilisateur à partir de ses textes. Ci-dessous l'analyse de la personnalité d'Oprah Winfrey basée sur ses tweets.



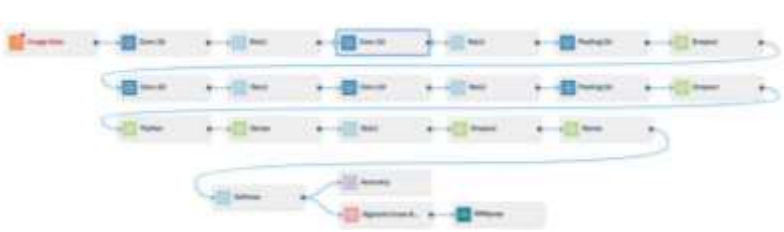
- **Language Translator** pour la traduction de textes et documents. Il supporte 30 langues dont les grandes langues européennes et asiatiques.
- **Speech to Text** et **Text to Speech** pour compléter des chatbots en entrée et en sortie pour en faire des assistants personnels. L'ensemble supporte neuf langues.
- **Tone Analyzer** analyse les émotions dans les textes comme dans les mails entrants de clients, dans les conversations des réseaux sociaux ou dans les chatbots. Cela renvoie une information sur le type d'émotion : colère, peur, …

Il nous reste aussi **Visual Recognition** qui est destiné à toutes les applications de reconnaissance de l'image. Il est très utilisé dans le diagnostic dans l'imagerie médicale. Cette solution peut être exploitée dans des endroits inattendus comme avec la startup française 3D-minded, et son application mobile "Le Croqueur" qui identifie les chocolats de grands chocolatiers à partir d'une photo, une sorte de Shazam du chocolat.

Les outils de création d'application d'AI comprennent :

- **Watson Studio**, l'environnement de travail qui permet à Watson de gérer ses projets d'IA. Il permet notamment d'accéder à des modèles de machine learning partagés dans la communauté des développeurs et sur quelques marchés (RH, marketing, supply chain…). Il gère tout le cycle d'un projet d'IA avec la récupération des données, la création des modèles de machine learning et de deep learning (avec un « network modeler » qui permet de créer des réseaux de neurones de manière graphique, *ci-dessous*), la préparation des données, l'entraînement des modèles, leur déploiement dans le cloud puis leur gestion en production. L'environnement gère aussi la distribution des traitements, notamment sur des serveurs et des GPU de type Nvidia. L'outil fonctionne en mode texte, en mode commande ainsi qu'en mode interactif avec des menus tout prêts pour lancer les batches.

  Watson Studio est destiné aux experts métiers, aux data scientists et aux développeurs[1629]. Il peut opérer les traitements dans le cloud, sur son propre micro-ordinateur ou sur des serveurs internes à l'entreprise.



- **AutoAI** est le module d'AutoML de Watson qui sert à identifier le bon modèle de machine learning en fonction des données à exploiter et des problèmes à résoudre. Ce module a été lancé en juin 2019. En pratique, AutoAI sert à automatiser la préparation et le prétraitement des données, au développement de modèles de machine learning et à l'optimisation de ses hyperparamètres.

[1629] Voir Introducing IBM Watson Studio, mars 2018.

Il est complété par IBM Neural Networks Synthesis (**NeuNets**) qui aide à optimiser les réseaux de neurones côté vitesse ou précision[1630].

- **Watson Knowledge Catalog** est une sorte de gestionnaire des données qui vont ensuite alimenter les autres briques de Watson. C'est un *repository* des données récupérées qui peuvent être aussi bien quantitatives (tabulaires, bases de données) que textuelles. L'outil permet de faire des *business analytics* sur ces données. Il fait partie de Watson Studio.

- **Watson Explorer** permet de créer des solutions personnalisées d'analyse de données structurées et non structurées. Sachant que depuis début 2018, la nouvelle version de **Cognos Analytics** supporte maintenant des fonctions cognitives diverses comme le dialogue en langage naturel.
- **Watson Machine Learning** recouvre les services en cloud de machine learning et deep learning qui sont exploitables dans Watson Studio. Les frameworks supportés comprennent TensorFlow, Spark ML, scikit-learn, XGBoost, Keras, PyTorch et le modeleur d'IBM SPSS[1631]. Watson Machine Learning gère le déploiement des solutions dans le cloud d'IBM.
- **Watson OpenScale** comprend les outils collaboratifs de mise en production, de monitoring et de diagnostic des applications d'IA réalisées avec les briques logicielles de Watson. Il comprend des outils de détection d'anomalies et de biais dans les modèles de machine learning créés ainsi que des outils d'explication des résultats. Le tout s'opère à partir d'une console de monitoring.
- **Content Hub** est en fait l'offre de CMS (content management system) d'IBM qui peut très bien tourner sans IA.
- **Watson Services for Core ML** permet de déployer des modèles entraînés de machine et deep learning pour les faire tourner sur iOS. Ce module était lancé en mars 2018.

IBM s'est aussi lancé dans les AIOps avec **IBM Watson AIOps** qui s'est renforcé en 2020 avec l'acquisition d'**Instana** (2015, USA/Allemagne, $57M).

Idem pour la RPA avec celle de **WDG Automation** (2014, Brésil) en 2020 qui est spécialisée dans l'automatisation de processus clients via le traitement du langage.

Watson est aussi décliné sur quelques marchés spécifiques au niveau applicatif avec **Watson Oncology** (cancérologie), **Watson Agro 4.0** (pour l'agriculture), **Watson Cybersecurity**, **Watson Virtual Agent** (des chatbots préconfigurés s'appuyant sur IBM Watson Assistant) ainsi que **Watson Maximo** (pour l'industrie 4.0 et les objets connectés).

Créer une application Watson revient donc souvent à créer du code, du contenu et à réaliser un travail d'intégration pour créer un agent conversationnel intelligent ou des applications de machine learning ou de deep learning. Dans des approches verticales, il faut définir des scénarios de dialogues assez précis et avoir sous la main beaucoup de données exploitables, aussi bien structurées que non structurées.

[1630] Voir IBM adds AutoAI to Watson Studio to speed up data prep so data scientists can focus on models par Larry Dignan, juin 2019.

[1631] Voir Deep Learning as a Service Now in IBM Watson Studio, mars 2018.

D'où l'importance pour IBM d'avoir un écosystème de partenaires solutions à même de couvrir les besoins de divers marchés verticaux. Pour ce faire, IBM a lancé un programme partenaire assez classique qui comprend l'accès aux APIs, à une communauté, un programme d'accélération de trois mois et un catalogue de solutions pour promouvoir les partenaires. À ce jour, l'écosystème d'IBM Watson comprend environ plusieurs centaines de sociétés dont un bon nombre de startups.

Le programme d'accélération porte surtout sur l'accompagnement technique mais donne aussi l'opportunité de présenter son offre pour récupérer un part du fonds d'investissement de $100M créé pour l'occasion.

En plus de son écosystème, IBM développe l'activité de services pour prendre en main de bout en bout les projets de ses grands clients. Alors que l'équipe d'origine de Watson ne faisait que quelques personnes, elle comprendrait maintenant plus de 10 000 personnes dans le monde, principalement des consultants, avant-vente et développeurs, dont environ 800 en France, y compris, un centre d'avant-vente et de support situé à Montpellier.

IBM a aussi ouvert un centre de recherche dédié à l'IOT et Watson à Munich associé à un investissement de $200M, probablement pluriannuel, et décliné Watson sur l'IOT avec des outils notamment dédiés à l'analytics et au machine learning.

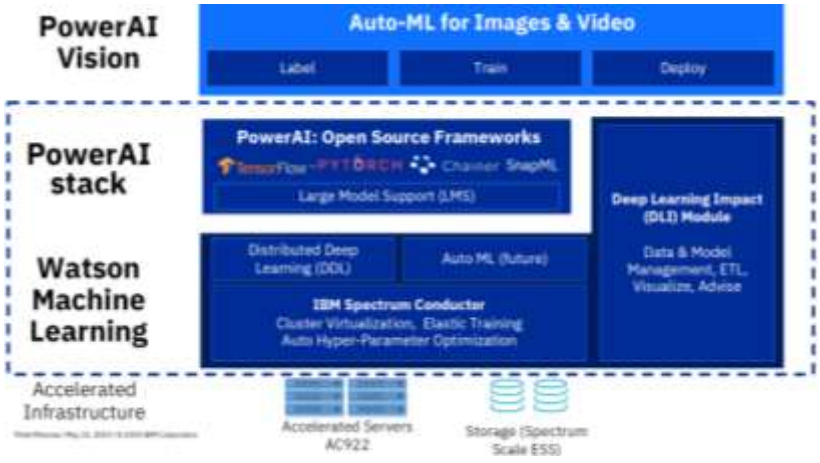
Ces quelques milliers de personnes allouées à Watson sont un bon début mais encore peu au regard des plus de 200 000 collaborateurs d'IBM Services. La migration d'IBM vers un business "cognitif" suffisamment différencié des autres sociétés de services globales dans le monde est une course contre la montre. Et ces dernières ne se laisseront probablement pas faire, même si elles auront probablement quelque temps de retard à l'allumage et du mal à recruter (ou former, si on peut rêver) les talents en machine et deep learning.

Quid des tarifs de Watson ? Il serait fourni à coup de licence logicielle d'un prix supérieur au million de dollars, mais avec un tarif plus proche de ceux du cloud pour les partenaires. Le tout est complété par les inévitables prestations de service associées.

Comment IBM se positionne-t-il par rapport à l'éventail des solutions du marché dans l'IA ? La comparaison est des plus difficiles. Il faut faire la part des choses entre briques prêtes à l'emploi, outils de développement propres à IBM, frameworks tiers, outils de distribution des applications dans le cloud d'IBM. L'offre la plus voisine semble être celle de Microsoft. En Inde, WiPro propose aussi sa plateforme **Holmes** (*Heuristics and Ontology-based Learning Machines and Experiential Systems*) qui concurrence directement Watson[1632].

IBM a aussi lancé **Power AI**, qui est le pendant « infrastructure ouverte » de Watson.

En gros, c'est une offre matérielle et cloud générique capable de faire tourner des applications d'IA développées avec les outils du marché tels que PyTorch et TensorFlow complétées par les outils de Watson sans que j'aie bien compris à quel niveau se faisait l'intégration.



[1632] Holmes est une plateforme d'IA hybride. C'est une boîte à outils voisine de celle d'IBM Watson avec de quoi créer un agent conversationnel, de la vision artificielle, du machine learning, à piloter drones et robots.

Enfin, **IBM Spectrum Discover** gère les couches basses de la consolidation et de la préparation des données pour les applications de machine learning. Il se connecte aux environnements de stockage hétérogènes tels que ceux de Dell EMC, Isilon, NetApp, Amazon S3 et Ceph et comprend des fonctions de classification, de labellisation des données et d'extraction de métadonnées. Il détecte aussi divers types d'informations personnelles confidentielles et données sensibles, ce qui sera utile pour gérer la conformité au RGPD. Il est complété par **IBM Spectrum Conductor** qui est la plateforme de déploiement des solutions créées avec Spectrum Discover.

Lancé en 2018 et étendu en 2020, IBM entretient un centre d'IA sur le plateau de Saclay dans le cadre d'**AIDA** (Artificial Intelligence for digital automation), une initiative qui regroupe des entreprises françaises, IBM et Bpifrance, ce dernier finançant un tiers des 33M€ du budget. L'idée est de développer une plateforme d'apprentissage à destination des entreprises sous la forme de coinnovation entre entreprises privées et laboratoires de recherche, un peu comme le font les IRT et les 3IA[1633].

Les données

IBM définit dans sa communication ce qu'est un bon projet pour Watson :

- Il doit traiter un **gros volume de données**. Cela fait sens !
- La solution doit permettre de **répondre rapidement** aux questions des utilisateurs, dans cette logique d'agent conversationnel fonctionnant en mode questions/réponses.
- La **variété des questions** traitées doit être grande grâce à une large palette de compréhension. Le système doit pouvoir traiter en profondeur les questions posées.
- Watson doit être en mesure **d'évaluer la validité des réponses**, avec un indice de confiance, comme il le faisait dans Jeopardy.

Les projets doivent être longs à conclure et à mener avec les grandes entreprises surtout si elles doivent mettre de l'ordre dans leurs données, comme ce fut le cas avec les projets de systèmes experts dans les années 1980.

Ils ont probablement également des clients dans les secteurs militaires et du renseignement US qui ne donnent pas lieu à de la communication marketing. Finalement, les références sont maintenant bien plus nombreuses avec les partenaires éditeurs de logiciels qu'avec IBM en direct.

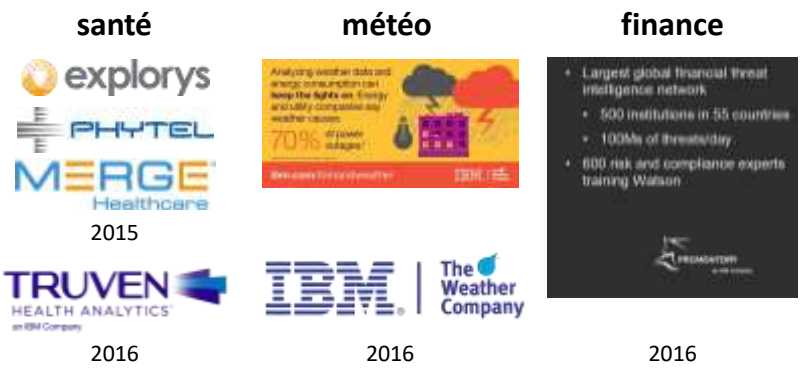
IBM est très dispersé verticalement mais avec un discours assez creux par secteur et relativement peu de références clients hors USA. Les outils marketing et la communication presse d'IBM répète le même discours générique sur Watson avec un zeste de vertical. Quant aux déploiements annoncés, il est toujours bien difficile d'évaluer s'ils sont sortis de la phase pilote.

Pour renforcer sa présence dans certains marchés verticaux, IBM a fait quelques acquisitions clés ou annonces :

- Avec **The Weather Company** pour $2B en 2016, afin d'alimenter Watson avec des données météorologiques destinées à diverses applications comme, par exemple, pour identifier les risques météorologiques dans la définition de primes d'assurances dans l'immobilier ou encore pour prévoir le trafic de clients dans le retail.
- Avec **Promontory** en 2016 et ses outils de gestion de risques et de conformité permettant d'étoffer son offre dans la finance.

[1633] Voir Intelligence artificielle : IBM lance en France un projet de R&D mondial par Ingrid Vergara, 2020. On notera que cette initiative est un moyen de contourner le fait qu'en 2019, le projet de création d'un 3IA sur le plateau de Saclay avait été retoqué par le jury international qui les avait sélectionnés. Les gagnants étaient les 3IA de Paris, Toulouse, Nice/Sophia Antipolis et Grenoble.

- Dans la santé, avec **Explorys** (2009, USA, $15M) avec sa plateforme de cloud dans la santé, **Phytel** (1996, USA, $22,5M) et sa solution de suivi de prise de traitements, **Merge Healhcare** (1987, USA) et ses outils de gestion d'imagerie médicale et **Truven Health Analytics** (2012, USA) et ses outils d'analytics.



  En 2018, IBM semblait dégraisser les effectifs de certaines de ces acquisitions. Mais c'est un processus assez traditionnel après des acquisitions de startups et celles-là n'y ont pas échappé. C'est en partie lié au fait qu'IBM a laissé tomber le marché de la cancérologie. Ils investissent encore dans la simulation moléculaire, avec leur outil maison, RXN for Chemistry. Manque de bol, c'est plutôt Google Deepmind qui faisait l'actualité en 2020 avec la nouvelle édition d'Alphafold qui repoussait les limites dans la simulation du repliement de protéines.

- IBM lançait début 2020 une offre d'optimisation de contenus publicitaires dans le cadre du lancement de l'Advertising Accelerator with Watson, qui n'est pas un accélérateur de startups[1634].

IBM a investi au moins $7B en acquisitions dans l'IA, bien plus que Google ne l'a fait. En plus des startups évoquées ci-dessous, il a notamment absorbé en 2014 la startup **Cognea** (2013, USA), créatrice d'un agent conversationnel, **AlchemyAPI** (2005, USA, $2M), une startup de deep learning d'analyse de textes et d'images, de reconnaissance de visages, de labellisation automatique d'images acquise en 2015, et **IRIS Analytics** (2007, Allemagne), une startup d'analyse temps-réel dédiée à la détection de fraudes aux moyens de paiement, s'appuyant sur du machine learning.

Le matériel

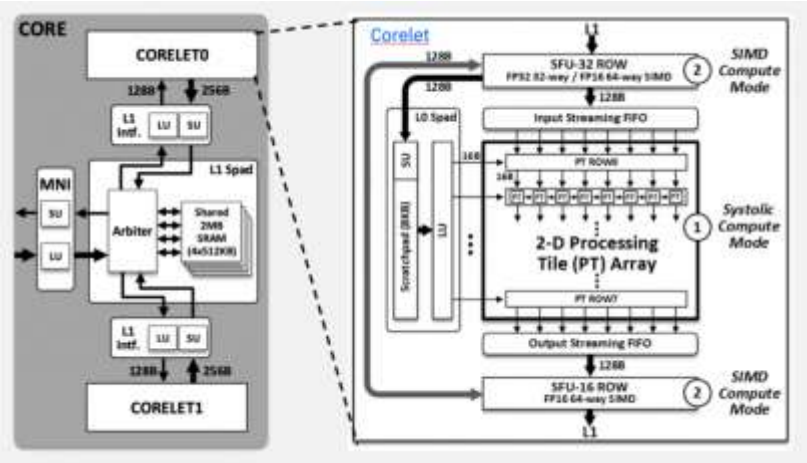
Nous avons vu dans les parties concernant les processeurs neuromorphiques et quantiques qu'IBM était un acteur intéressant avec d'un côté ses processeurs TrueNorth et de l'autre, ses premières expériences d'ordinateur quantique qui sont disponibles dans le cloud.

Tout ceci permet à IBM de conserver un peu d'avance dans sa capacité à produire des calculateurs de haute performance. Mais l'industrialisation à grande échelle est le bât qui blesse chez IBM. Pour que ces investissements soient rentables, il leur faudra générer du volume et trouver des débouchés pour ces composants. En effet, pour ce genre de technologie, rien ne dit que l'intégration verticale soit la meilleure approche. Surtout si la concurrence se structure de manière horizontale comme Intel le fait avec succès sur le marché des PC et des serveurs depuis 35 ans.

Fin 2017, IBM lançait ses serveurs Power9 censés être optimisés pour les applications d'IA. Leur processeur Power9 est un CPU assez classique mais très puissant. Ses AC922 Power Systems comprennent deux CPU Power9 avec 2 To de mémoire et quatre GPU Nvidia V100 reliés en NVLink. Tout ceci permet d'accélérer les transferts de données entre les composants, un point très important pour l'entraînement de modèles de machine learning. Celà étant, cette architecture n'est pas idéale pour entraîner des modèles de deep learning. On retrouve cette architecture dans le supercalculateur IBM Summit déployé en 2018.

[1634] Voir IBM Launches Advertising Accelerator with Watson, janvier 2020.

IBM Research planche sur la conception de composants d'entraînement et d'inférences d'IA plus génériques avec un module fonctionnel comprenant deux cœurs de tenseur fonctionnant en nombres flottants 16 bits et partageant la même mémoire SRAM. Le principal intérêt de l'ensemble est d'améliorer l'efficacité énergétique du processeur[1635]. Reste à intégrer cela dans un processeur commercial ou déployé dans les infrastructures de cloud d'IBM.



Google

Google est aujourd'hui considéré comme le GAFA le plus en pointe côté IA, que ce soit par la dominance de TensorFlow chez les développeurs, ou par le fait que ses services d'IA ont la plus grosse base installée au monde avec Search et Android, et via leur recherche fondamentale et appliquée hyperactive.

On en vient à oublier que Google était au départ un projet de création d'une vaste IA mondiale, démarré par un simple moteur de recherche. Ce serait devenu un véritable projet de société alimenté par un solutionnisme débridé[1636].

Presque tous les services de Google font appel à de l'IA : dans Google Search en général, dans la recherche d'images similaires de Google Search ou Google Photos, dans l'antispam de Gmail[1637], dans l'assistant personnel Google Assistant et les devices Google Home, en liaison avec ses objets connectés pour la maison, dans ses Google Car et dans Android. Sans compter la filiale Waymo d'Alphabet, la maison mère de Google.

S'y ajoute pour épater la galerie les performances multiples de leur filiale **DeepMind**[1638], et notamment AlphaGo ou ses IA qui jouent à StarCraft et autres jeux multi-joueurs complexes. Et aussi l'ensemble de leur infrastructure de cloud avec tous les outils associés qui complètent les briques d'IA.

[1635] Voir A 3.0 TFLOPS 0.62V Scalable Processor Core for High Compute Utilization AI Training and Inference par Jinwook Oh et al, 2020.

[1636] Voir la critique du modèle dans «La Vie après Google» ou l'espoir d'un monde nouveau, octobre 2018 au sujet de « Life after Google » de l'économiste George Gilder, qui entrevoit un monde décentralisé sauce Blockchain remplaçant le monde centralisé de Google.

[1637] L'IA d'antispam de gmail générerait seulement 0,05% d'erreurs. Elle exploite un système de deep learning réparti sur 16 000 CPU avec plus d'un milliard de connexions entre neurones.

[1638] Elle serait un gouffre financier, mais Google peut se permettre ce genre d'investissement long terme. Voir DeepMind's Losses and the Future of Artificial Intelligence, août 2019.

Il est relativement facile de différencier Google d'IBM.

Les deux maîtrisent des technologies logicielles somme toute assez voisines. La différence principale réside dans la manière de les mettre entre les mains des clients. Les deux ont des plateformes logicielles et cloud mises à disposition des développeurs et des startups.

Tandis qu'IBM fonctionne en mode projet et service avec les grandes entreprises, Google propose des services à plus de deux milliards d'internautes. Cela lui permet d'accumuler d'énormes volumes de données qu'il utilise pour entraîner ses propres IA, ce qu'IBM n'arrive à faire que sur certains marchés verticaux et via quelques acquisitions ciblées. Google a-t-il intérêt à copier IBM ? Pas vraiment. La profitabilité et la croissance de Google sont excellentes alors qu'IBM a une profitabilité de société de services et est en décroissance.

L'histoire

L'actualité abonde depuis 2014 d'acquisitions médiatisées de startups de l'IA par les grands acteurs du numérique, Google en tête. Cela alimente quelques fantasmes sur leurs avancées qui sont quelque peu enjolivées. Google aurait, selon certains, acquis tout ce qui existerait de bien comme compétences dans l'IA. C'est évidemment une vue de l'esprit.

Oui, Google a fait bien plus d'acquisitions dans le domaine de l'IA que les autres grands du numérique, mais rappelons-nous le côté très artisanal de ce secteur. Ce n'est pas parce que vous achetez quelques verreries de luxe que vous êtes le seul à savoir fabriquer des verres de luxe ! L'artisanat est très souvent un marché très fragmenté. On peut le constater au regard des effectifs des startups acquises. Ils sont en général très réduits, comme ils l'étaient d'ailleurs pour les acquisitions par Facebook de startups telles qu'Instagram, Whatsapp ou Oculus Rift, qui n'avaient par ailleurs aucun rapport avec l'IA.

L'acquisition la plus médiatisée de Google dans l'IA fut celle de **DeepMind** (2010, UK) en 2014 pour un montant record dans ce secteur de $625M. Et surtout, pour à peine une cinquantaine de personnes dont une douzaine de chercheurs en machine learning. DeepMind s'est depuis surtout fait remarquer en étant à l'origine des différentes moutures d'AlphaGo ou de réseaux de neurones comme PathNet. Il mêne une recherche dans pas mal de domaines allant d'applications de l'IA dans la santé à la quête du Graal de l'AGI (Artificial General Intelligence).

Google avait auparavant mis la main sur la société de reconnaissance vocale **SayNow** (2005, USA, $7,5M) en 2011 puis sur **Viewdle** (2006, USA, $12M) et **PittPatt** (2004, USA) en 2012, qui faisaient tous les deux de la reconnaissance faciale et de mouvements. En 2013, ils mettaient la main sur le spécialiste des réseaux neuronaux **DNNresearch** (2012, Canada), et embauchaient ainsi le canadien Geoff Hinton, considéré comme le père du deep learning. Mais il n'y travaille visiblement pas à temps plein. Google perdait en tout cas Ian Goodfellow, le père des GANs, en 2019, parti chez Apple.

Ont suivi **Dark Blue Labs** (2014, UK) et **Vision Factory** (2014, UK), deux sociétés d'Oxford qui n'ont pas levé de fonds. S'y ajoutèrent le spécialiste de la traduction automatique **Quest Visual** (2009, USA), et celui de la reconnaissance de mouvements **Flutter** (2010, USA, $1,4M) qui a probablement enrichi l'offre logicielle de Dropcam, une startup de caméras de surveillance qui est dans le giron de Nest, une filiale d'Alphabet.

Depuis 2017, cette belle frénésie ininterrompue d'acquisitions dans l'IA semble s'être calmée. Dans l'IA, on ne compte que celle d'**AIMatter** (2016, Biélorussie, $2M) avec sa plateforme de deep learning mobile de reconnaissance d'images et puis celle d'**Onward** (2015, USA, $120K), une plateforme de création de chatbot de service client.

Google a aussi acquis **Moodstocks** (2008, France, $500K) qui proposait une solution mobile de reconnaissance d'images, fournie sous la forme d'APIs et d'un SDK multi-plateforme. Cela semble être une acquihire.

L'année 2014 avait vu Google/Alphabet acquérir une belle brochette de startups dans la robotique avec **Schaft** (robot humanoïde et bras articulé, Japonais), **Industrial Perception** (robots industriels, spécialisé dans la vision 3D), **Redwood Robotics** (bras robotisés, issue du SRI et acquise un an après sa création), **Meka Robotics** (aussi dans les bras robotisés, qui avait contribué à la création de Redwood Robotics), **Holomni** (roues robotisées), **Bot & Dolly** (bras articulés à mouvements très souples servant aux tournages de cinéma) et **Autofuss** (encore des bras articulés).

Il y avait surtout **Boston Dynamics**, connu pour ses robots médiatisés doués de capacité de marche à quatre puis deux jambes mais que Google a cédé à **Softbank Robotics** en juillet 2017, cédé par ce dernier au coréen **Hyundai** en 2020. Google souhaitait aussi céder **Schaft** à Softbank Robotics la même année mais cela n'a pas aboutit et Google a tout bonnement stoppé l'activité de Schaft sans le céder. Bref, la stratégie robotique « attrape tout » de Google était bien à prendre avec des pincettes.

Que deviennent toutes ces acquisitions ? Ce qui relève de l'imagerie et du langage s'est retrouvé dans les services de Google. La robotique n'a débouché sur aucune application commerciale. Google l'utilise dans ses véhicules utilisés pour cartographier les rues. Et Google ne cherche pas à concurrencer les leaders de robots industriels (ABB, Fanuc, etc)[1639].

Google fait des progrès réguliers dans le traitement des images, comme avec **PlaNet** qui identifie à quel endroit ont été prises des photos d'extérieur ou pour compter les calories dans des **photos de plats cuisinés**. Google utilise aussi beaucoup d'IA sensorielle pour faire évoluer les fonctions de conduite automatique de ses Google Car.

L'IA googlelienne a connu un sursaut de médiatisation début 2016 avec la victoire de la solution AlphaGo contre le champion du monde coréen de Go, **Lee Sedol**, construite par une équipe d'une vingtaine de personnes de sa filiale DeepMind (vidéo de la première partie)[1640]. Ces victoires ont été présentées comme des étapes importantes des progrès de l'IA, faisant écho à la victoire de Deep Blue aux échecs contre Gary Kasparov en 1997.

[1639] Voir Google robotics: A review - The Robot Report, octobre 2017, et Google Has Made a Mess of Robotics de Mark Bergen et Joshua Brustein, octobre 2017, qui racontent les déboires de Google dans la robotique.

[1640] Le champion prenait sa retraite fin 2019. Voir Former Go champion beaten by DeepMind retires after declaring AI invincible par James Vincent, novembre 2019.

La différence ? Le jeu de Go est plus difficile à simuler car la combinatoire de jeu est bien plus grande qu'aux échecs. AlphaGo ne peut donc pas compter que sur la force brute. Il doit combiner plusieurs méthodes pour être efficace : éliminer des options de jeu inutiles via le "Monte Carlo Tree Search" ou MCTS et exploiter une base de jeux permettant d'identifier des tactiques gagnantes. Il réalise ensuite un apprentissage supervisé à base de deep learning en exploitant 150 000 parties connues. Il choisit ses coups avec un réseau de neurones convolutif. Il fait de l'apprentissage par renforcement en jouant contre lui-même. Et l'ensemble exploite la puissance machine de GPU et de TPU que nous avons vue dans la partie consacrée aux processeurs neuromorphiques. En 2017, AlphaGo Zero faisait progresser la discipline en s'entraînant par lui-même et en étant plus économe en ressources. Nous l'avons déjà évoqué.

Dans "Artificial Intelligence and the Singularity" publié en 2016, Piero Scaruffi se faisait un malin plaisir de relativiser cette victoire (*ci-dessous*) en rappelant la consommation d'énergie du système par rapport aux 20W du cerveau humain[1641] !

La performance a été documentée dans un article publié dans la revue Nature en janvier 2016[1642]. Un peu vexés, les coréens ont d'emblée lancé un plan de financement public de 765M€ dans l'IA sur cinq ans avec les géants comme Samsung, LG, Hyundai et SK Telecom[1643]. C'était suivi par un nouveau plan de $2B lancé en mai 2018[1644].

En 2015, Matthew Lai développait **DeepChess**, un système de deep learning avec renforcement qui gagnait aux échecs en apprenant lui-même à optimiser son jeu en moins de 72 heures sur un simple PC. Il était recruté par DeepMind début 2016 et il a contribué aux évolutions d'AlphaGo à partir de ce moment-là !

Tout cela faisait en tout cas une excellente publicité pour DeepMind dont les solutions de machine learning ont heureusement d'autres applications comme la **curation de médias**, même si elles font moins parler d'elles car elles ressemblent de près à ce qu'IBM fait déjà dans la santé avec Watson. Leur **DQN** est un réseau neuronal profond doté de capacités d'auto-apprentissage et **DeepMind Health** et donne lieu à une collaboration avec la NHS britannique dans l'application Streams de détection de blessures aux reins dans les urgences. En novembre 2018, Alphabet annonçait regrouper ses activités disparates dans la santé sous l'ombrelle de Google Health.

[1641] On pourrait ajouter que depuis plus de 30 ans, n'importe quel tableur gagnerait haut la main toute partie contre les champions du monde du calcul mental !

[1642] Voir Mastering the game of Go with deep neural networks and tree search, janvier 2016 (37 pages).

[1643] Voir Intelligence artificielle : la Corée (vexée d'avoir perdu au jeu de go) annonce un plan de 765 M€, de Julien Bergounhoux, mars 2016. En mai 2017, une version améliorée d'Alpha Go battait le champion chinois Ke Jie.

[1644] Voir South Korea Aims High on AI, Pumps $2 Billion Into R&D, mai 2018.

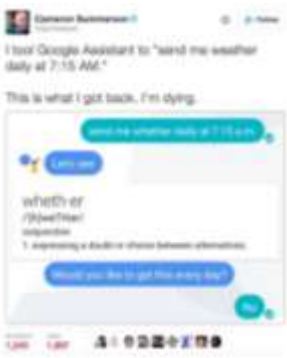
Parfois, la communication de Google DeepMind en fait un peu trop comme avec Psychlab, un toolkit qui évalue les capacités cognitives d'une IA à base d'agents (UNREAL = UNsupervised REinforcement and Auxiliary Learning) en s'appuyant sur les mêmes modèles psychophysiques que pour l'évaluation des capacités cognitives humaines à se déplacer dans l'espace pour atteindre un objectif[1645]. Il évalue ici les IA d'apprentissage par renforcement dans le contexte de la détection visuelle de formes. L'évaluation montre que l'apprentissage par renforcement a une capacité à planifier ses actions sur le long terme (vidéo). Tout cela est sympathique mais c'est de l'IA très très étroite. L'agent UNREAL s'appuie sur un réseau convolutif (CNN) couplé à un réseau à mémoire (LSTM).

DeepMind déploie ses efforts dans de nombreux domaines de la santé : pour détecter les rechutes de vétérans aux USA à partir d'un historique de 700 000 dossiers médicaux, pour détecter le cancer du sein, pour détecter des pathologies de l'œil, pour améliorer le traitement du cancer du cou et de la tête. Hors santé, ils ont développé l'application Adaptive Battery qui prédit l'application que l'on va utiliser et améliore les performances de la batterie et WaveNet, déjà évoqué pour le text-to-speech. Parfois, la filiale de Google échoue dans ses tentatives, comme celle de réussir des examens de mathématique de lycées de 16 ans[1646].

Google investit aussi pas mal pour créer une IA de confiance et éthique, tant que faire ce peut. Cela se voit par exemple via leur création du **Federated Learning** qui préserve la protection des données privées servant à entraîner des modèles de deep learning centralisés. En 2019, ils rendaient Google Assistant en partie autonome sur Android, évitant des transferts intempestifs de la voix des utilisateurs sur Internet. Ils ont évidemment leur propre charte d'éthique, assez classique. Cela ne les empêche cependant pas de se prendre parfois les pieds dans le tapis comme avec le renvoi intempestif pas bien justifié en décembre 2020 de Timnit Gebru, une chercheuse de couleur sur l'éthique de l'IA qui copilotait l'équipe éthique de Google AI. Cela a déclenché l'ire de milliers de salariés de Google et de chercheurs extérieurs à Google[1647].

Les logiciels

Google utilise l'IA pour enrichir ses propres offres grand public, que ce soit autour de son moteur de recherche multifonctions ou de business plus périphériques d'Alphabet (santé, IoT, automobile). On la retrouve aussi dans Google Assistant et Google Home, ces agents conversationnels pilotables à la voix et au clavier.



C'est aussi un grand fournisseur de plateformes de développement, le plus souvent en open source, et dans l'embarqué ou en cloud.

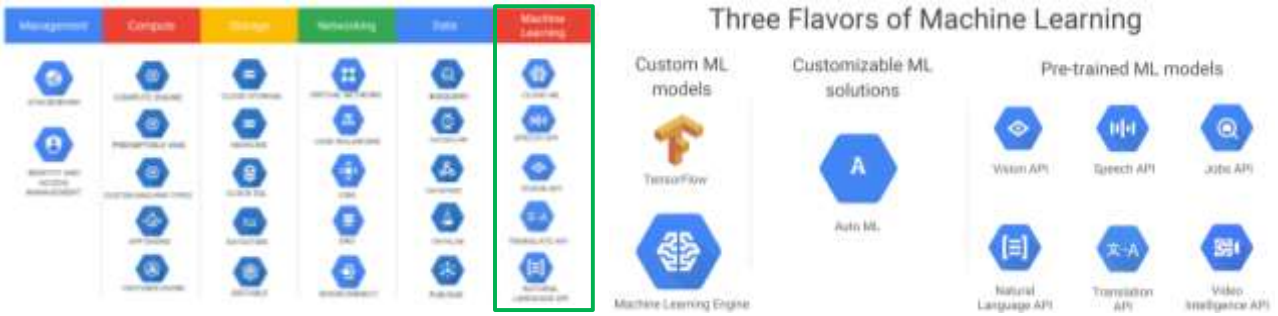
Il publie régulièrement de nombreuses **APIs de services d'IA dans le cloud** pour les développeurs (Google Cloud Machine Learning Engine). Google est aussi à l'origine de la bibliothèque de machine et deep learning **TensorFlow** qui est très couramment utilisée par les startups de l'IA. Comme IBM, nombre de ses services couvrent le traitement du langage, y compris la traduction, ainsi que la vision artificielle.

[1645] Voir Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents, janvier 2018 (28 pages).

[1646] Voir DeepMind : l'IA championne de Google se plante à un examen de mathématiques simple par Pierrick Labbe, 2019.

[1647] L'objet de la contention était la publication d'un *review paper* sur les biais potentiels des grands modèles de NLP comme BERT et GPT-3. Comme ceux-ci sont alimentés par de grandes masses de contenus captés sans trop de filtrage sur Internet, ils peuvent contribuer à véhiculer si ce n'est amplifier des biais discriminateurs de la société. John Dean, le boss, considérait que ce papier ne correspondait pas aux standards de la société. Voir Google's Dismissal Of A Top AI Researcher Shows How Little Progress Has Been Made In Listening To Discrimination Concerns par Mark Murphy, décembre 2020 et We read the paper that forced Timnit Gebru out of Google. Here's what it says par Karen Hao, décembre 2020.

Cette panoplie couvre une majeure partie des besoins de créateurs d'applications à base d'IA. Le machine learning est proposé sous plusieurs formes : à bas niveau avec les outils de développement de TensorFlow, ensuite avec ses outils d'**AutoML** qui proposent automatiquement des modèles de machine learning en fonction de la structure des données et des besoins et enfin, avec des modèles pré-entraînés pour le traitement du langage et de la vision.



Google AI Platform Pipelines est un service de gestion des pipelines de machine learning à partir de composants et modèles TensorFlow Extended (TFX). Le service **ML Metadata Management** fournit pour sa part un registre des actions et modèles. Google propose aussi les **OR-Tools**, une bibliothèque logicielle pour faire de la programmation par contrainte, gérer des optimisations linéaires et des algorithmes à base de graphes et flux.

Lancé en novembre 2019, **Contact Center AI** (CCAI) sert à déployer des agents virtuels de relation clients. En septembre 2020, il était complété par **Dialogflow CX**, une suite de création de chatbots pour centres d'appels avec laquelle on peut définir le parcours utilisateur de manière visuelle (vidéo).

Mais Google regorge de développeurs et d'équipes projets en tout genre. Nous avons par exemple les projets **Google Brain**[1648] lancés en 2011 par une petite équipe de chercheurs qui comprend Jeff Dean, Greg Corrado, Andrew Ng et Geoff Hinton depuis 2013.

Cette équipe est à l'origine de systèmes de chiffrement évolutifs publiés en octobre 2016[1649] et d'un étonnant programme d'amélioration d'images pixellisées publié en en février 2017, exploitant des images de 8x8 pixels[1650] pour augmenter leur résolution à 32x32 pixels.

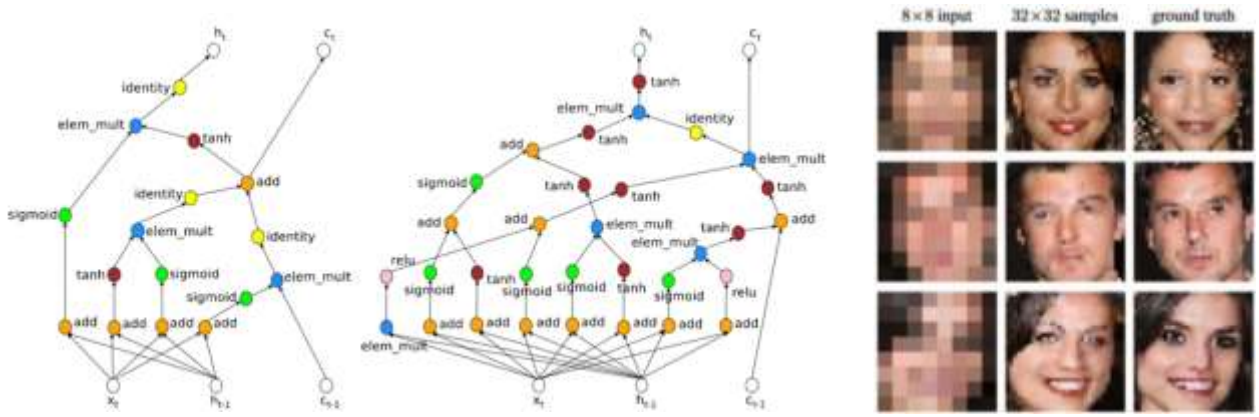
L'image du milieu est celle que l'IA de Google reconstitue à partir de celle de gauche. C'est impressionnant. Mais attention au fait que les images de départ semblent être des versions basse résolution de la base d'entraînement.

[1648] Google Brain est distinct de Deep Mind qui est basé au Royaume Uni, reste relativement indépendant après son acquisition en 2014.

[1649] Qui là aussi a beaucoup fait fantasmer avec « l'IA qui créé son propre langage que les hommes ne comprennent pas ». Voir Google taught artificial intelligence to encrypt messages on its own de Dave Gershgorn, octobre 2016.

[1650] Voir Pixel Recursive Super Resolution, février 2017. Qui rappelle le scénario du film « No way out » avec Kevin Costner, sorti en 1987. Il faut préciser que le système est entraîné avec les images de la dernière colonne.

En mai 2017, ils publiaient aussi d'étonnants travaux montrant comment ils pouvaient utiliser le machine learning pour améliorer l'architecture d'un réseau de neurones qui est d'habitude créée manuellement (*illustrée ci-dessous à gauche*[1651]).



Les développeurs de Google sont à l'origine d'avancées régulières dans les réseaux de neurones de reconnaissance d'image. C'est le cas de **Facenet** qui améliore les techniques de reconnaissance de visages, entraîné sur 260 millions d'images et efficace à 86% en 2016. La méthode ? Une variante de réseau de neurones convolutif[1652] (*ci-dessus à droite*). L'équipe de Google Brain est aussi à l'origine d'améliorations diverses de Google Translate[1653].

Lors de la conférence Google I/O de mai 2018, Google présentait quelques autres nouveautés en matière d'IA, surtout du traitement du langage : des évolutions de Google Image dont une fonction de colorisation, reprenant des algorithmes génératifs déjà bien connus, l'auto-remplissage d'emails avec la fonction Smart Compose (vidéo) et Google Duplex (vidéo) pour faire prendre des rendez-vous au téléphone par un assistant vocal.

Autant la première démonstration est plausible et s'appuie sur des modèles probabilistes de réseaux de neurones à mémoire (LSTM ou variante), autant la seconde était sujette à caution tant que l'on n'a pas pu la tester dans une grande variété de cas. Mais elle est maintenant en production, Google reconnaissant cependant faire parfois appel à des humains.

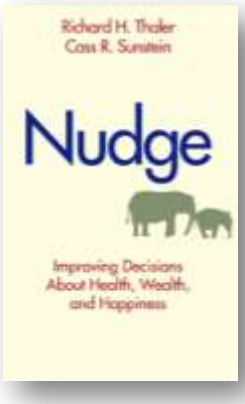
On peut aussi citer l'extension Chrome **DeepBreath** lancée en 2017 et qui prévient l'utilisateur en train de rédiger un email un peu incendiaire pour lui recommander de réfléchir avant de l'envoyer. Ce type de fonctionnalité était évoquée de manière prémonitoire dans l'ouvrage Nudge de Richard Thaler et Cass Sunstein en 2008 (voir *ci-dessous*). Et puis la fonctionnalité **Smart Compose** qui remplit automatiquement la fin des phrases lorsque l'on rédige ses emails sous Gmail. Et la fonctionnalité est disponible en français depuis 2019.

Android contient depuis sa version 9 (Pie) des « actions » et « slices », des fonctions pour les développeurs qui permettent de prévoir la prochaine action de l'utilisateur (vidéo). Mais ces prévisions sont à faire par les développeurs. Android Pie ne fournit que les briques logicielles et interfaces de programmation permettant de les intégrer dans Android. C'est à se demander ce qu'il reste du libre arbitre comme l'évoque à juste titre Gaspard Koenig !

[1651] Voir Using Machine Learning to Explore Neural Network Architecture, mai 2017.

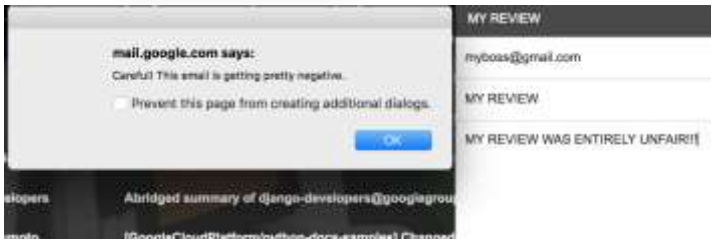
[1652] Voir FaceNet: A Unified Embedding for Face Recognition and Clustering, juin 2015. L'un des trois auteurs, James Philbin, a quitté Google en 2015. Il est depuis le directeur de la vision artificielle de **Zoox** (2014, $290M), une startup ultra-bien financée qui veut devenir un opérateur de service de véhicules autonomes.

[1653] Voir Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016.

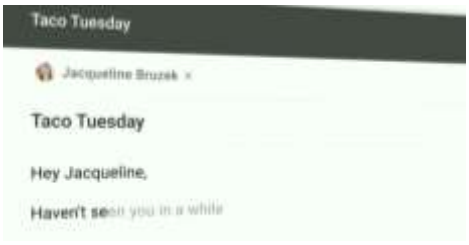


We propose a Civility Check that can accurately tell whether the email you're about to send is angry and caution you, "WARNING: THIS APPEARS TO BE AN UNCIVIL EMAIL. DO YOU REALLY AND TRULY WANT TO SEND IT?" (Software already exists to detect foul language. What we are proposing is more subtle, because it is easy to send a really awful email message that does not contain any four-letter words.) A stronger version, which people could choose or which might be the default, would say, "WARNING: THIS APPEARS TO BE AN UNCIVIL EMAIL. THIS WILL NOT BE SENT UNLESS YOU ASK TO RESEND IN TWENTY-FOUR HOURS." With the stronger version, you might be able to bypass the delay with some work (by inputting, say, your Social Security number and your grandfather's birth date, or maybe by solving some irritating math problem!).*

Nudge 2008



Google DeepBreath Chrome plugin 2017



Gmail Smart Compose 2018

Les données

Il va sans dire que la puissance de Google vient de la quantité astronomique de données qu'ils accumulent sur les faits et gestes des internautes et même dans le monde physique de millions d'utilisateurs. En gros, Google sait ce que l'on recherche (Search), où l'on est et où l'on va (Android, Maps), quels moyens de transport on utilise (Maps, Android), ce que l'on échange avec les autres (Gmail), le temps que l'on passe sur tel et tel écran, et un peu plus rarement, ce que l'on cherche et regarde à la TV (Google TV).

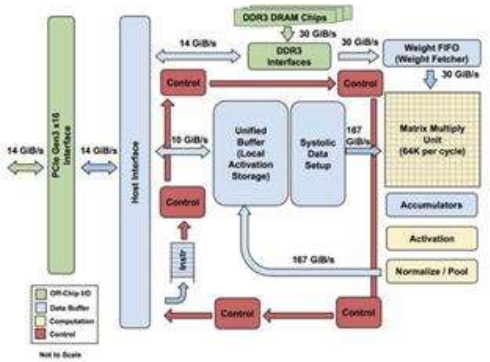
Ils ont la compétence pour stocker, indexer et gérer ces données dans le cloud. A partir de là, ils peuvent tester tout un tas d'idées sans grandes limites ! Mais il ne faut pas non plus exagérer leur puissance. Ils n'accèdent pas aux données métier des entreprises et en particulier à celles des infrastructures. Ils n'ont pas accès aux données bancaires, d'assurances, de la distribution ou dans la santé, même si dans ce dernier cas, ils fournissent les ressources techniques de calcul en cloud à des chercheurs en génomique à divers laboratoires de recherche[1654].

Le matériel

Google a développé ses propres processeurs neuromorphiques en 2015/2016, les TPU[1655]. Ce sont des processeurs programmables capables de gérer des applications de deep learning. Ils ont été utilisés par AlphaGo de DeepMind et gèrent bien d'autres briques d'AI de Google. Ils sont aussi proposés sous forme de ressources dans le Cloud de Google. En 2020, Google annonçait sa quatrième génération de TPU. Nous les avons déjà décrits dans la rubrique sur les processeurs neuromorphiques.

[1654] Voir Economic and social impacts of Google Cloud, par Deloitte, 2018 (109 pages).

[1655] Voir First in-depth look at Google's TPU archtecture, de Nicole Hemsoth, avril 2017.



Google n'a pas l'air de prévoir la commercialisation de ces processeurs et serveurs. C'est une technologie fabriquée en volume en ASIC destinée à équiper leurs propres datacenters. Mais comme Google est le premier consommateur de serveurs au monde, ils ont les économies d'échelle qui le permettent.

Pour ce qui est des ordinateurs quantiques, Google a commencé par tester avec la NASA les ordinateurs à recuit quantique du Canadien **D-Wave**. Google met aussi au point ses propres ordinateurs quantiques à qubits supraconducteurs – effet Josephson. Ils annonçaient en mars 2018 avoir atteint 72 qubits, mais sans les avoir démontrés dans une machine opérationnelle. En octobre 2019, ils annonçaient avoir atteint la suprématie quantique avec le processeur Sycamore de 53 qubits.

Ils ont sinon déjà créé deux frameworks de développement quantique : **Cirq** pour le développement à bas niveau et **OpenFermion** pour le développement d'applications dans le domaine de la simulation chimique.

Google est aussi concepteur d'objets connectés intégrant des logiciels d'IA, qu'il s'agisse de **Google Home**, des smartphones **Pixel** ou des futurs véhicules de **Waymo**, qui seront probablement fabriqués par des constructeurs automobiles tiers, Google fournissant l'électronique, les logiciels (sous Android) et les solutions en cloud associées[1656].

Amazon

Amazon est un acteur clé de l'IA dans les entreprises via son offre intégrée de cloud sous la bannière des **Amazon Web Services** (AWS). Amazon est le leader mondial des services génériques en cloud, utilisé largement par les entreprises tout comme par les startups. Il devance Microsoft Azure qui est le second sur ce marché.

amazon

offre	forces	faiblesses
Lex, Rekognition, Polly Alexa (pf NLP) Echo (device) Spark (big data distribué) AWS (cloud services) DSSTNE (deep learning) Kiva (robots)	services grand public données utilisateurs utilisateur intensif de ML dans ses propres services leader mondial du cloud logistique médias et contenus	faible différenciation logicielle dans l'IA pas habitué à évangéliser les développeurs comme IBM, Google et Microsoft

[1656] A noter que le Directeur de l'Engineering de Waymo, la filiale pour véhicules autonomes de Google, est le Français Sacha Arnoud, issu de l'X.

Amazon est comme tous les GAFAMI un acquéreur de startups régulier, mais pas très actif du côté de l'IA. On peut noter celle du spécialiste des robots d'entrepôts **Kiva** (2003, USA, $18M) en 2012 pour $775M, d'**Ivona** (2004, Pologne), spécialiste du text to speech, acquis en 2013, d'**Angel.ai** (2015, USA, $8M) en septembre 2016, créateur d'un chatbot généraliste qui a certainement dû les aider à améliorer Alex et de **harvest.ai** (2014, USA, $2,74M) détection de failles de sécurité, acquis en 2017.

Kiva, robots d'entrepôts, acquis en 2012 pour $775M

Safaba, traduction automatique, acquis en 2015

Orbeus, recherche d'images, acquise en 2016

Angel.ai, chatbot généraliste acquis en 2016

harvest.ai, détection de failles de sécurité, acquis en 2017

Ivona, acquis en 2013

Les logiciels

Vu du grand public, Amazon est aussi présent dans l'IA via les nombreuses déclinaisons d'**Amazon Echo** et le service en cloud de dialogue en langage naturel **Alexa** qui l'équipe et qui est très largement utilisé par l'écosystème des objets connectés.

Il est devenu le standard par défaut des objets à commande vocale. Il est supporté par un nombre incalculable d'objets connectés : radioréveils, robots, set-top-boxes, lampes connectées, copycats de l'Amazon Echo, routeurs, dans l'électroménager, dans les aspirateurs robots et même dans les véhicules connectés. Il devance de ce point de vue Google et son Assistant.

Il y aurait plus de 100 000 applications pour Alexa dont 65901 pour les USA et 1567 pour la France [1657]. Google a rattrapé du terrain face à Amazon avec son Assistant et ses hauts parleurs connectés mais Amazon domine encore largement ce marché. Il représentait en 2019 plus des deux tiers des ventes de haut-parleurs connectés aux USA[1658].

L'offre d'APIs d'AI pour les développeurs d'applications en cloud comprend, outre Alexa :

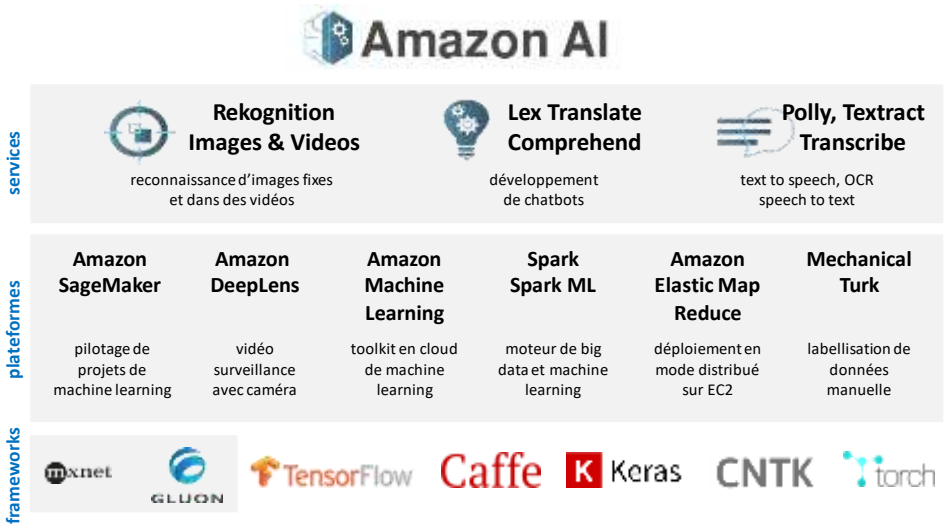
- **Amazon Rekognition**, une fonctionnalité d'analyse d'images fixes et de vidéos à base de deep learning qui permet d'identifier des objets, de les tagger, d'éliminer des contenus illicites, d'analyser les expressions dans les visages et de les reconnaître.
- **Amazon Polly** est une solution de text-to-speech réaliste lancée fin 2016 (vidéo et conférence technique) avec un choix de 47 voix dans 25 langues.
- **Amazon Transcribe** réalise la fonction inverse, le speech to text. Il est par exemple exploitable pour créer des transcriptions texte des appels de clients et les injecter dans des systèmes de

[1657] Voir Amazon Alexa Has 100k Skills But Momentum Slows Globally. Here is the Breakdown by Country par Bret Kinsella, octobre 2019.

[1658] Voir Alexa devices maintain 70% market share in U.S. according to survey par Greg Sterling, août 2019.

CRM. S'y ajoute **Amazon Transcribe Medical** une déclinaison destinée au monde médical de leur solution de speech-to-text, lancée en décembre 2019[1659].

- **Amazon Textract AI** est un service d'OCR en cloud qui reconnaît les formats de documents avec tableaux, factures et autres formulaires. L'outil génère des données structurées avec les données extraites. Cela peut servir par exemple à exploiter des scans de factures ou de notes de frais.
- **Amazon Lex** et **Comprehend** sont les moteurs de gestion de conversations d'Alexa.
- **Amazon Translate**, est un moteur de traduction issu de la startup américaine **Safaba Translation Solutions** (2009, USA) acquise en 2015.



Du côté des couches basses, nous avons :

- **Amazon SageMaker** est un environnement complet de gestion du cycle de gestion des projets de machine learning. Il lance toutes les briques intermédiaires (vidéo). On charge les données, entraîne des modèles, les teste puis les met en production. Les services de SageMaker se pilotent à partir d'un environnement de développement (notebook) en mode web dénommé **Jupyter**. En décembre 2019, Amazon lançait Sagemaker Studio, une version web de son outil de développement[1660]. Elle comprend maintenant une fonctionnalité d'AutoML baptisée AutoPilot ainsi que SageMaker Pipelines pour la gestion de worflow d'envergure de machine learning. SageMaker Feature Store gère de manière centralisée ses modèles de machine learning pour les réutiliser d'un projet à l'autre. SageMaker Datawrangler se consacre au processus d'acquisition, ingestion et préparation des données d'entraînement de modèles de machine learning. SageMaker Clarify sert enfin à détecter les biais dans les données d'entraînement.
- **DeepLens** est une solution complète de reconnaissance d'images intégrant une caméra et les logiciels de traitement des vidéos qu'elle génère. La caméra utilise un processeur Atom tournant sous Linux avec une puissance de 100 Gflops, 8 Go de mémoire et 32 Go de stockage. Elle capte les photos avec 4 mpixels en MJPEG et les vidéos Full HD en H.264. La caméra a probablement été conçue par les équipes de la startup **Blink** (2009, USA, $5,8M), acquise en décembre 2017. La partie serveur s'appuie sur Amazon Rekognition Video.
- **Amazon Machine Learning** consolide les outils de création et exécution de modèles de machine learning.
- **Spark** pour la distribution de traitements sur serveurs, logiciel de la fondation Apache et **SparkML** ou Spark MLlib, une bibliothèque qui permet de distribuer des traitements de machine learning ainsi que **BigDL**, une bibliothèque de deep learning. Tous ces logiciels sont open source ! Les entreprises payent les ressources en cloud pour les héberger.

[1659] Voir Amazon debuts automatic speech recognition service, Amazon Transcribe Medical par Sarah Perez, décembre 2019.

[1660] Voir AWS launches SageMaker Studio, a web-based IDE for machine learning par Frederic Lardinois, décembre 2019.

- **DSSTNE** (Deep Scalable Sparse Tensor Network Engine, ou « destiny ») qui permet de créer des modèles de machine learning et de deep learning faciles à déployer sur GPU… Nvidia en général.
- L'interface de programmation **Gluon** lancée par Amazon et Microsoft en 2017 qui supporte diverses plateformes comme MXNet d'Amazon et le CNTK de Microsoft[1661]. Cela ressemble à un concurrent du framework Keras qui permet de décrire simplement les couches d'un réseau de neurones et leur entraînement. Gluon ne semble pas générer de traction chez les développeurs deux ans après son lancement.
- **Amazon Elastic MapReduce**, qui permet de déployer des applications Spark et Hadoop sur plusieurs serveurs, notamment pour le traitement de gros volumes de données.

Au niveau des frameworks :

- **Mxnet** est un framework complet de la même catégorie que PyTorch et TensorFlow, et optimisé pour fonctionner sur le cloud d'Amazon.
- **Gluon** est un framework de plus haut niveau au-dessus de Mxnet. C'est en quelque sorte l'équivalent de Keras pour Tensorflow.

Les données

Accessoirement, Amazon est le leader mondial du commerce en ligne et la part qu'il représente dans ce marché est en croissance, surtout aux USA où il captait 37,7% du marché en 2019. Mais 57% en intégrant les ventes des tiers sur sa place de marché. Cela représente un peu plus de 5% des ventes de détail aux USA.

Amazon possède le plus gros inventaire de produits dans son catalogue, qui est estimé à 350 millions de références, comprenant les offres intégrées dans sa place de marché. En conséquence de quoi, comme Google, il dispose d'un beau pactole de données pour analyser les comportements des Internautes dans leur casquette de consommateurs. Il dispose aussi de données sur la consommation culturelle via ses services Prime Video, ses tablettes Kindle et sa box TV Fire, surtout aux USA.

Amazon est donc un gros utilisateur et de longue date de techniques de machine learning pour optimiser tout son processus de vente et de logistique. Il les utilise pour planifier la demande et gérer au plus près les stocks, pour définir les prix, les offres de livraison, pour la recommandation de produits, la détection de fraudes et de contrefaçons, pas toujours parfaite d'ailleurs. Cela explique probablement pourquoi Amazon est l'un des plus gros recruteurs aux USA de spécialistes de machine learning[1662].

Enfin, rappelons qu'Amazon propose depuis 2005 **Mechanical Turk**, un service d'intermédiation avec des « ouvriers de la donnée » payés quelques dollars de l'heure qui labellisent des données servant aux systèmes de machine learning[1663]. Les tâches les plus courantes : labelliser des images et transcription d'audio. Le service aurait plusieurs centaines de milliers de participants.

Le matériel

Depuis fin 2017, les serveurs EC2 d'AWS permettent l'accès à des GPU **Nvidia** V100, avec des grappes de 8 équivalentes aux serveurs Nvidia DGX1. Ceux-ci sont notamment dédiés à l'entraînement de modèles de deep learning. AWS propose aussi depuis fin 2019 des serveurs d'inférences utilisant les chipsets d'Habana, maintenant intégré à Intel, ainsi que des serveurs équipés de FPGA Xilinx, de Nvidia A100 et de CPU AMD ou à base de noyeaux arm.

[1661] Voir AWS and Microsoft announce Gluon, making deep learning accessible to all developers par Microsoft, octobre 2017.

[1662] Voir IBM (IBM) Creating Shareholder Value with AI? Not so Elementary, My Dear Watson de Jefferies, juillet 2017.

[1663] Voir Untold History of AI: How Amazon's Mechanical Turkers Got Squeezed Inside the Machine par Oscar Schwartz, avril 2019.

Amazon serait en train de développer ses propres chipsets embarqués pour faire fonctionner Alexa en s'appuyant sur les compétences de la startup **Anapurna** (2011, USA) acquise en 2015. En décembre 2020, Amazon lançait **Trainium**, un chipset d'entraînement de modèles de machine learning pour ses serveurs, sans que ses caractéristiques soient précisées. Il succédait au chipset **Inferentia**, lancé fin 2019.

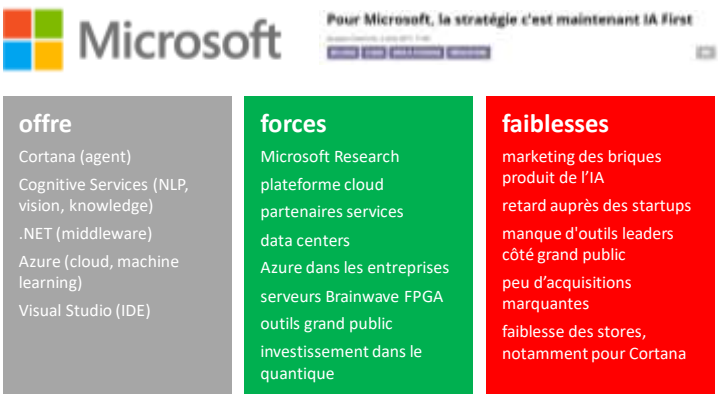
En décembre 2020, Amazon lançait également **AWS Panorama**, un serveur local d'interprétation de vidéos de caméras de surveillance utilisable notamment pour le contrôle qualité en usine. Le boîtier est durci (IP62). Il est associé à un SDK, se connecte d'un côté à des caméras en réseau TCP/IP détectables via le protocole ouvert du consortium **ONVIF** et de l'autre, à du matériel intégrant des chipsets dédiés de vision artificielle comme les Jetson de Nvidia. Mais il est déjà équipé d'un processeur AGX Xavier de Nvidia.

Par ailleurs, du côté logistique, ils sont évidemment friands de robotique comme nous l'avons déjà vu. Les robots qu'ils ont acquis chez **Kiva** servent à déplacer des palettes dans les entrepôts et Amazon est en train d'évaluer des robots de préparation des colis de A à Z.

Microsoft

Microsoft est un acteur de poids dans les infrastructures informatiques des entreprises et son arsenal logiciel dans l'IA est équivalent sur de nombreux points avec celui de ses concurrents tels qu'IBM et Google.

De manière assez classique, son offre couvre le traitement du langage, de la vision artificielle, des données et la gestion des connaissances, ainsi que tout ce qui permet de gérer des modèles de machine et deep learning. L'éditeur est plus discret dans sa communication autour de l'IA et son marketing produit est moins efficace, tout du moins comparativement à IBM et Google. Par contre, l'éditeur a mis toute la voilure sous le vent de l'IA depuis 2017.



Microsoft met aussi le paquet pour promouvoir une "IA responsible et éthique", comme dans l'ouvrage The Future Computed, Artificial Intelligence and its role in Society, 2018 (152 pages) créé par l'équipe des juristes et de relations institutionnelles de Microsoft. En complément, le Guide de survie de l'IA publié par Microsoft France en 2018 est un catalogue d'études de cas par marchés verticaux réalisées avec des partenaires services et logiciels.

En France, ils ont lancé une école de l'IA de formation sur 7 mois de jeunes et personnes au chômage, opérée par Simplon.co qui est maintenant déclinée en région comme à Lyon.

A Station F, leur IA Factory accélère des startups de l'IA. Les personnes formées passent ensuite 12 mois en alternance chez des partenaires de Microsoft. Cela concerne une douzaine de personnes par session.

En mars 2018, Microsoft annonçait devenir partenaire de l'**Institut PRAIRIE** (PaRis Artificial Intelligence Research InstitutE) créé conjointement par le CNRS, Inria, et PSL et des entreprises privées, dont Amazon, Google, Facebook, Criteo, Faurecia, Naver Labs, Nokia Bell Labs, PSA, Suez et Valeo. Cela sera donc un temple de la recherche partenariale en IA. Cela complète le laboratoire de recherche conjoint entre Microsoft et Inria situé à Palaiseau et lancé en 2005.

Les logiciels

Microsoft a ceci de commun avec IBM qu'il entretient depuis des décennies de grandes équipes de recherche fondamentale et particulièrement investies dans les différents champs de l'IA.

Créé en 1991, **Microsoft Research** occupe plus de 1000 chercheurs répartis dans le monde, y compris en France, dans un laboratoire commun monté à Orsay avec Inria. La principale équipe européenne est située à Cambridge au Royaume-Uni. L'équipe de Microsoft Research a été associée aux équipes technologiques en charge de l'IA en 2017. L'ensemble comprend 8000 collaborateurs.

Microsoft Research emploie un nombre record de prix Nobel et de scientifiques ayant gagné la médaille Fields. Cela n'en fait pas pour autant les initiateurs de business significatifs pour Microsoft. Tout au plus sont-ils à l'origine de nombreuses innovations incrémentales qui ont alimenté les produits phares de l'éditeur. Le correcteur orthographique qui souligne les mots dans Word était ainsi sorti de ces laboratoires en 1995. Cela permet de relativiser le rôle de la recherche pour dominer une industrie. Apple qui n'a pas formellement de laboratoire de recherche domine ainsi le secteur du mobile, en compagnie de Google ! Chez Google, la frontière entre recherche et développement est plus floue. Microsoft et Google sont en fait très proches dans leur équilibre de R&D.

Les activités de Microsoft Research dans le machine learning sont imposantes avec plusieurs dizaines d'équipes projets impliquées. Les projets comprennent les grands classiques qui portent sur l'amélioration de la reconnaissance de la parole et des images et notamment le tagging automatique de vidéos[1664].

Et puis, en vrac, un agent conversationnel détectant des troubles psychiatriques (**DiPsy**), un outil de reconnaissance de chiens originaire de Chine qui fonctionne à l'échelle individuelle, pas à celui de la race (**Dog Recognition**) et un outil de tri de pièces de monnaie pour les réfractaires aux Blockchains (**Numiscan**).

Les équipes de Microsoft Research sont à l'origine d'avancées comme le système de dialogue en langage naturel **Cortana**. Comme nombre de technologies d'IA proviennent de MSR chez Microsoft, l'éditeur se retrouve à mettre systématiquement en avant les travaux de ses chercheurs, parfois un peu trop au détriment des équipes produit classiques.

Microsoft qui est maintenant résolument tourné vers le cloud fait tout de même quelques acquisitions de startups pour accélérer son "time to market" dans l'IA ou dans la périphérie de l'IA. Les équipes de recherche fondamentale travaillent en effet sur des domaines où le risque est plus scientifique et technique que marché tandis que les startups sont censées œuvrer sur un risque marché.

Le risque est même parfois émotionnel et dans l'image, comme l'a montré le robot conversationnel **Tay** qui s'est mis à tenir des propos nazis et a été débranché. Tay était sorti de Microsoft Research et ses propos relevaient d'un apprentissage supervisé non filtré ! Tay a été remplacé en avril 2017 par un autre chatbot au doux nom de **Zo** qui est intégré dans la messagerie instantanée **Kik**. Zo est une version anglaise d'un chabot chinois de Microsoft dénommé **Xiaoice**. Mais Kik n'est pas très trendy chez les Internautes !

Les acquisitions dans les startups de l'IA sont relativement peu nombreuses chez Microsoft. En 2015, l'éditeur avait mis la main sur **Prismatic** (2010, USA, $15M), un agrégateur de news s'appuyant sur du machine learning, ainsi que **Double Labs** (2013, USA), une application Android de notification elle aussi basée sur du machine learning.

En 2016, c'était au tour de **Revolution Analytics** (2007, USA ? $38,7M), qui faisait de l'analyse prédictive s'appuyant sur le langage open source R, acquise en 2016. Un moyen de s'attirer un écosystème de développeurs ! Toujours en 2016, **Swiftkey** (200, UK, $21,6M), un logiciel de clavier virtuel mobile qui s'appuierait lui aussi sur du machine learning.

[1664] Voir un panaché des applications de Microsoft Research dans les produits de Microsoft dans From search to translation, AI research is improving Microsoft products par Jennifer Langston, décembre 2019.

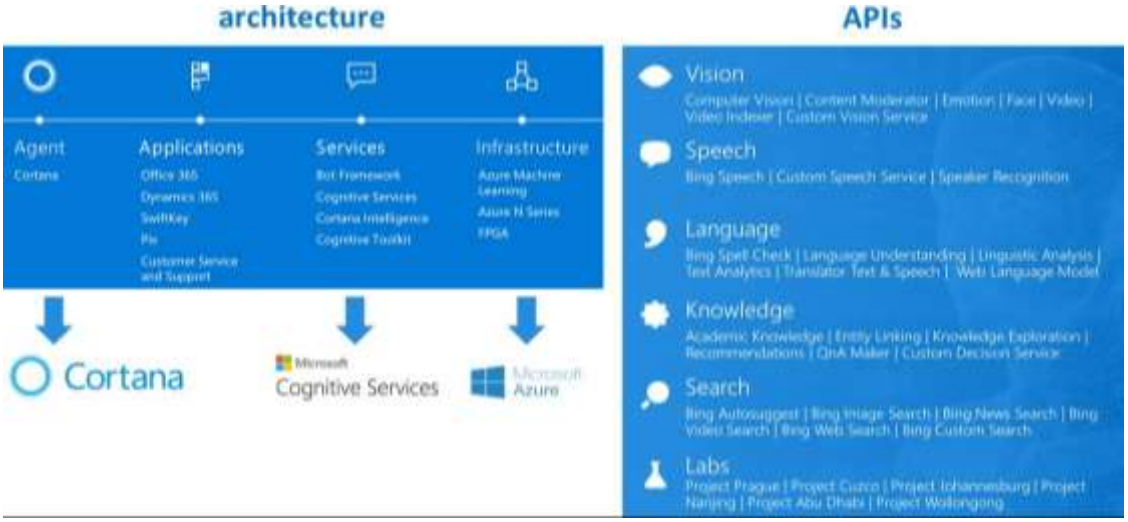
En 2017, Microsoft faisait l'acquisition de **Genee** (2014, $1,45M) un gestionnaire d'agenda virtuel et du Canadien **Maluuba** (2011, $8,2M), un spécialiste du deep learning appliqué au traitement du langage qui travaille sur l'AGI (Artificial General Intelligence), dont les équipes se sont fait remarquer en faisant gagner leur solution au Pac-Man avec leur technique Hybrid Reward Architecture[1665] (vidéo).

En juin 2018, **Bonsai** (2014, USA, $13,6M) apportait dans la corbeille son BRAIN (Basic Recurrent Artificial Intelligence Network), un système de machine learning avec un haut niveau d'abstraction, destiné surtout aux systèmes embarqués, et doté de capacités d'apprentissage par renforcement, comme pour l'entraînement de robots à saisir des pièces. La startup avait été créée par deux anciens ingénieurs de l'éditeur, Mark Hammond et Keen Browne. Retour au bercail !

En septembre 2018, c'était enfin le tour de **Lobe** (USA) de passer sous le giron de Microsoft. La startup a développé un système de création visuelle de modèles de deep learning à base de glisser-déplacer.

On peut enfin citer l'acquisition fin 2018 de **Xoxco**, dans le développement de chatbots[1666] et en juillet 2019, l'investissement de $1B dans **OpenAI** qui est passé par la même occasion du statut d'ONG à startup classique.

Il n'empêche que l'éditeur a bien compris les enjeux de l'IA et cherche à se positionner comme fournisseur de plateforme d'IA pour les développeurs, avec le "Conversation As a Platform" et le "Microsoft Bot Framework", qui rappellent dans leur structure l'offre des APIs d'**IBM Watson**.



L'architecture d'IA de Microsoft est un maquis de briques logicielles avec un mix d'outils de développement, de déploiement dans le cloud et d'applicatifs horizontaux ou verticaux. Le « branding » de l'ensemble laisse à désirer et est encore plus instable que chez IBM qui a au moins « Watson » comme cri de ralliement.

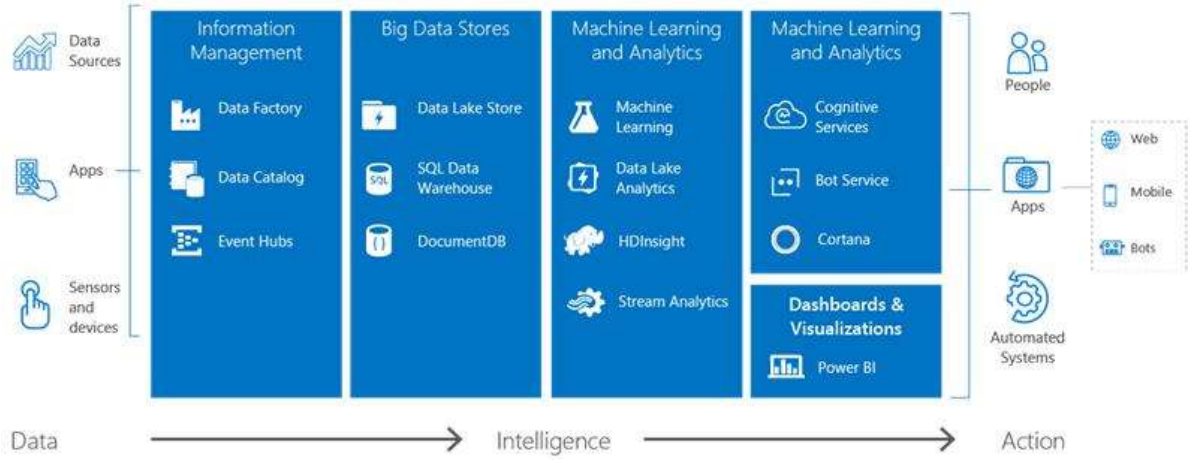
- Des **outils de création** d'applications avec l'IDE Visual Studio (Integrated Development Environment) et ses Code Tools for AI, et aussi l'outil **Azure Machine Learning Studio**, lancé en 2015, qui permet de créer ses modèles de machine learning et de les mettre en production. Annoncé en septembre 2017, Azure Machine Learning est maintenant découpé en trois modules avec Workbench pour la modélisation, Experimentation pour le test de modèles sur les infrastructures du cloud dont des GPU, et Model Management, pour les déploiements. Le tout avec une application native fonctionnant sous Windows et MacOS.

[1665] Voir Hybrid Reward Architecture for Reinforcement Learning, juin 2017, qui décrit une architecture d'apprentissage par renforcement avec des agents fonctionnant en parallèle.

[1666] Voir Microsoft to acquire Xoxco as focus on AI and bot developers continues par Ron Miller, novembre 2018.

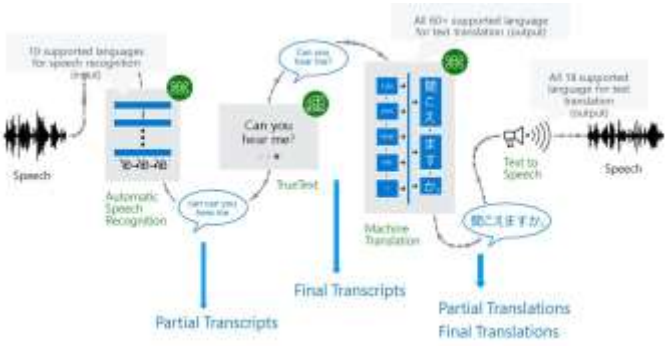
- Les **services cognitifs** qui comprennent plus d'une vingtaine d'APIs sensorielles qui font de la reconnaissance d'images, du traitement du langage naturel (NLP), de la gestion de connaissances et de la recherche. À bas niveau, Microsoft propose en open source son framework de "deep learning" **CNTK** (Computational Network Toolkit) depuis fin 2015. Les API de vision artificielle permettent par exemple de détecter les émotions dans les visages et d'estimer l'âge des personnes (*ci-dessous*).

- L'**agent Cortana**, bon à tout faire, répond à la voix et joue le rôle de traducteur[1667]. C'est un peu l'équivalent d'Amazon Alexa et de Google Assistant. Quelques bots ont été développés avec les APIs de Microsoft mais l'offre n'a rien à voir avec l'abondance qui sévit autour d'Amazon Alexa et plus récemment de Google Assistant.
- La plateforme en cloud **Azure** est ouverte et intègre nombre d'outils open source du marché avec notamment TensorFlow et Caffe (frameworks), et aussi Apache Spark (pour la distribution des traitements sur les serveurs) et Docker (pour le déploiement d'applications).



On trouve l'outil de création de chatbot **Microsoft LUIS** (Language Understanding Intelligent Service) lancé début 2017, l'outil de traduction **Microsoft Translator** et ses Translator Speech Translation API (*dont le processus est illustré ci-dessous*) ainsi que le **Microsoft Bot Framework** et le **BotBuilder** qui servent à créer son propre chatbot. Les **Visual Studio Tools** lancés en 2018 permettent de gérer des projets avec les principaux frameworks du marché tels le CNTK de Microsoft, Tensorflow, Keras et Caffe.

[1667] Voir A developer's guide to building AI applications – Create your first intelligent bot with Microsoft AI, 2018 (52 pages).

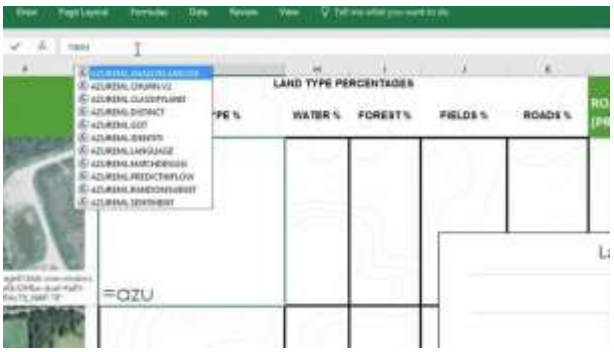


En mars 2018, Microsoft lançait **WinML**, un jeu d'APIs destinées à l'exécution sous Windows 10 de modèles de machine et deep learning déjà entraînés. Cela rappelle Core ML d'Apple qui est destiné à iOS. WinML doit intégrer la mise à jour de Windows 10 qui répond au doux nom de code Redstone 4. WinML est aussi calibré pour exploiter le chipset Intel Movidius qui est dédié à la reconnaissance d'images, notamment pour les webcams des laptops ou pour faire des recherches dans ses photos par une recherche naturelle sur leur contenu.

- **Azure AutoML** sert à identifier automatiquement son modèle de machine learning en fonction du jeu de données à utiliser. Il est adapté à la résolution de problèmes de classification, de prévisions et de régression.
- **ML.NET** est le jeu de bibliothèques open source de machine learning pour les développeurs .NET. Il supporte C# et F# sur Windows, Linux et même macOS. Il est complété par ML.NET Model Builder, une interface visuelle qui s'intègre avec AutoML[1668]. L'ensemble peut être étendu avec Tensorflow de Google et le format intermédiaire ONNX de description de réseaux de neurones.
- L'infrastructure en cloud **Azure** et les logiciels serveurs de Microsoft et tous les outils de supervision qui vont avec.
- Dans l'embarqué, l'**Azure IoT Edge SDK** permet de gérer des modèles de deep learning entraînés de reconnaissance d'image dans des caméras utilisant le chipset mobile Qualcomm QCS603 et le Vision SDK de ce dernier. Microsoft va aussi intégrer **Windows ML** dans une mise à jour de Windows 10, un run-time permettant d'exécuter des modèles entraînés d'IA dans ses applications, en exploitant le CPU et le GPU de l'ordinateur.
- Diverses **applications** qui intègrent des briques d'IA, comme Office 365, Dynamics 365 et l'application mobile de gestion de photos Pix. Le traitement du langage à base d'IA est disséminé dans Office, et depuis des années. En septembre 2017, Microsoft annonçait l'intégration de services de machine learning du cloud Azure dans Excel, qui se manifestent sous la forme de fonctions (*ci-dessous*). **Seeing AI** est une application qui décrit l'environnement vu d'un smartphone, très utile pour les personnes malvoyantes (vidéo). Des fonctions d'IA sont même intégrées dans Powerpoint[1669].

[1668] Voir Microsoft launches a drag-and-drop machine learning tool par Frederic Lardinois, mai 2019.

[1669] Voir Microsoft ajoute plus d'intelligence, et un coach, à PowerPoint par Mary Jo Foley, juin 2019.



Microsoft ajoutait des fonctionnalités d'IA dans **Dynamics**, les Dynamics 365 AI pour aider les commerciaux à prioriser leurs actions de prospection et de relance avec interrogation en langage naturelle, pour rendre les données clients accessibles via des chatbots et notamment via Cortana, puis « AI for Market Insights » qui « *vise à aider les équipes de marketing et de médias sociaux à accéder à des informations exploitables plus rapidement pour une réponse plus rapide aux besoins des clients* » ce qui ne veut strictement rien dire de précis[1670].

- On peut ajouter à tout cela **Microsoft Learn** qui est une offre de formation en ligne à tous ces outils. Et elle est gratuite. Vous pouvez par exemple vous former à la création d'un outil de classification automatique d'images[1671].

Lorsque Microsoft commercialise Azure et ses outils d'IA à une grande entreprise, c'est souvent rapidement présenté comme un partenariat stratégique.

C'est le cas de la data factory créée en 2019 avec **Intermarché**, ce qui est une manière déguisée de présentation de la vente de services en cloud du premier au second[1672]. Autre forme de partenariat, l'accélération conjointe de startups comme dans l'**AI for Green Energy**, créé avec **Schneider Electric** et intégré dans l'AI Factory de Microsoft à Station F. Il s'agit d'une nouvelle thématisation du choix des startups accompagnées dans l'IA qui opèrent dans la santé, l'environnement, l'énergie, les transports, les services financiers et l'agro-alimentaire.

Microsoft utilise d'autres effets de levier dans son implantation en régions ainsi que dans le lancement de programmes de formation à l'IA, sous-traités à **Simplon.co**. En 2018, Microsoft avait annoncé la création dans trois régions d'**experiences Lab** et de trois nouvelles Ecoles d'IA comme à Lyon[1673]. En 2019, les formations IA Microsoft / Simplon.co qui ciblent notamment des adultes en reconversion obtenait le label de la **Grande École du Numérique** (GEN). Cette formation dure 7 mois en alternance pour une vingtaine de participants par session. Microsoft est aussi partenaire d'**OpenClassrooms** pour proposer des cours en ligne sur l'IA[1674] et lançait enfin une « AI business school » en ligne avec l'**INSEAD**[1675], une formation en ligne gratuite destinée aux dirigeants.

Donc, dans la novlang marketing d'aujourd'hui, une vente devient un partenariat stratégique, le marketing se transforme en formation gratuite et une implantation en région se sous-traite.

[1670] Il faut consulter le document Applications de gestion – Notes de publication Octobre 2018, juillet 2018 (273 pages) pour comprendre qu'il s'agit de tableaux de bords associés à l'analyse de sentiments dans les réseaux sociaux.

[1671] Voir Classifier des images avec le Service Vision personnalisée de Microsoft.

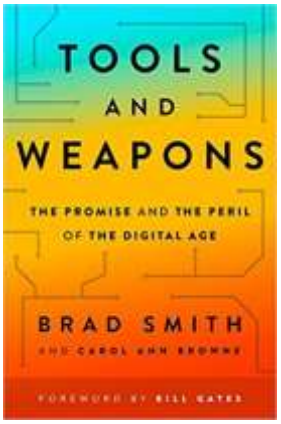
[1672] Voir Microsoft et Intermarché annoncent un partenariat stratégique pour développer le "mieux manger", renforcer la proximité et accélérer la transformation de la distribution avec tous les acteurs de l'écosystème alimentaire, par Juliette Français, Microsoft, juin 2019.

[1673] Voir Plan IA Microsoft - Microsoft France accompagne l'accélération de l'innovation au cœur des écosystèmes régionaux par Juliette Francaix, avril 2019 et L'école IA de Microsoft s'installe à Lyon chez Simplon par Véronique Arène, février 2019.

[1674] Voir Microsoft annonce un partenariat avec OpenClassrooms afin de repérer de nouveaux talents en IA par Arthur Vera, avril 2019.

[1675] Voir Microsoft et l'INSEAD lancent une formation unique sur le marché : l'AI Business School par Juliette Francaix, mars 2019.

Enfin, citons l'implication forte de Microsoft France dans l'association **Impact IA** qui vise à favoriser le développement d'IA éthique, en compagnie d'autres entreprises privées telles qu'Orange, des cabinets de conseil et quelques laboratoires de recherche publique. En juillet 2019, l'association présentée comme un « collectif » et un « do tank » fêtait son premier anniversaire avec 48 entreprises membres. L'association anime quatre groupes de travail : un observatoire, IA responsable, AI for good et formation/éducation[1676].



Tout ceci est en phase avec l'activisme de Microsoft aux USA et notamment de son Président Brad Smith qui milite pour réguler l'IA et contre, par exemple, l'abus de l'usage de la reconnaissance faciale[1677].

Les données

Microsoft dispose d'une activité grand public, certes pas aussi soutenue que celle de Google, mais qui lui permet d'avoir une forte expertise dans le cloud ainsi que dans la captation de données d'usages, à même d'alimenter ses outils de machine et deep learning. Il en va ainsi du moteur de recherche **Bing**, de **Skype**, de **MSN**, de la console de jeu **Xbox** et bien évidemment de **Windows**.

Le matériel

Comme Google et IBM, Microsoft a développé sa propre architecture serveur pour gérer des réseaux de neurones.

Elle s'appuie sur des processeurs développés en technologie FPGA. L'architecture s'inscrit dans le projet Brainwave dont les contours ont été dévoilés fin août 2017[1678] et qui s'appuient sur :

- Une **architecture** de serveurs massivement parallèle et distribuée associant CPU et FPGA.
- De **processeurs FPGA** fabriqués par Intel en technologie 14 nm, les Stratix 10 (ex Altera)[1679]. Ce sont des FPGA à mémoire qui stockent les paramètres des réseaux de neurones et évitent de faire appel à de la DRAM dans les serveurs, ce qui est bien plus rapide. Leur architecture de FPGA est dite « soft DNN » et donc reprogrammables, tandis que celle des TPU de Google n'est pas reprogrammable (« hard » DNN), ce qui apporte, pour faire simple, plus de flexibilité. L'architecture est optimisée à la fois pour des réseaux de neurones convolutifs (Convnets, pour la reconnaissance d'images) qui nécessitent de multiplier des matrices et des réseaux de neurones récurrents (RNN, LSTM et consorts, pour le traitement du langage). Cela leur apporte une plus grande flexibilité pour les déploiements à grande échelle dans leurs data-centers.
- Un **compilateur et un environnement d'exécution système** permettant de déployer des modèles du Microsoft Cognitive Toolkit tout comme de Google Tensorflow.

Cette architecture est déployée dans les datacenters de Microsoft Azure depuis 2016 mais l'éditeur est assez discret à son sujet.

[1676] Voir ce contre-point sur ce genre d'initiative: No, AI is not for social good par Jared Moore, novembre 2019.

[1677] Voir Microsoft president Brad Smith's book about tech is no love letter par Harry McCracken, septembre 2019.

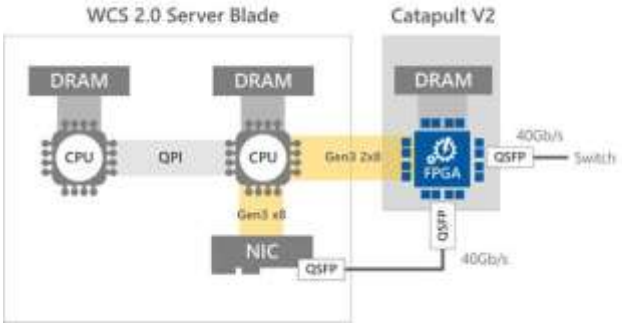
[1678] Voir Microsoft unveils Project Brainwave for real-time AI, août 2017.

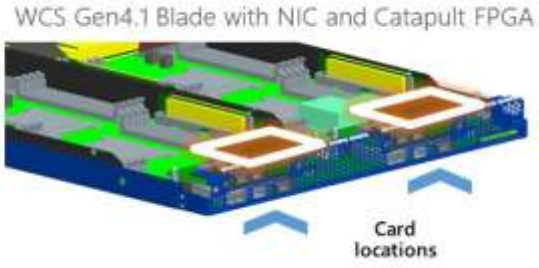
[1679] Microsoft avait produit ses premiers FPGA en 2011 dans ses serveurs Catapult V0 pour la gestion d'index de Bing. La V1 de Catapult sortait en 2012. En 2013, 1600 FPGA étaient mis en production. Capatupt V2 sortait en 2014 avec une architecture de bus optimisée pour faire communiquer les CPU avec les FPGA dans les serveurs, via un bus PCI à 64 Gbits/s (4 canaux).

En parallèle, Microsoft Research planche aussi sur des ordinateurs quantiques topologiques à base de fermions de Majorana[1680]. Mais aucun processeur opérationnel n'était encore sorti des laboratoires en 2021.

Intel/Altera Stratix 10

- 10 TFLOPS FP32
- HBM2 integrated
- Up to 1 GHz
- 14nm process
- 80 GFLOPS/W





Facebook

Le leader occidental des réseaux sociaux est avide d'IA à tous les étages pour améliorer l'ensemble de ses services que ce soit pour caser les bons contenus et les bonnes publicités aux bons endroits.

Les logiciels

Facebook est à l'origine de frameworks de machine learning et deep learning assez populaires chez les développeurs, derrière TensorFlow de Google : **PyTorch** qui est utilisé pour la recherche, **Caffe2** qui est utilisé en production et le langage de description de modèles de machine learning **ONNX**, co-créé avec Microsoft et Amazon (et sans Google).

Leurs solutions d'IA vont des fonctions de reconnaissance de photos dans leurs différents services mobiles à tout ce qui permet de mieux cibler les publicités pour optimiser les revenus en passant par le **Bot Framework** lancé en avril 2016 servant à la création de chatbots dans l'application Facebook Messenger.

facebook

offre	forces	faiblesses
API chatbots pour Messenger	services grand public	mal équipé pour accompagner les grandes entreprises
traitement d'images dans Facebook et Instagram	données utilisateurs	pas de véritable plateforme logicielle pour les entreprises
framework de machine learning PyTorch	Instagram et Whatsapp	image suite à l'affaire Cambridge Analytica
	Instant Messenger chatbot	
	laboratoire de recherche à Paris / Montréal / Yann Le Cun – Jérôme Pesenti	

En 2017, ils publiaient le logiciel open source **Prophet**, qui fait des prévisions de données temporelles à base de machine learning exploitant des régressions.

En mai 2019, Facebook mettait aussi en open source trois nouveaux outils : **PyText NLP**, un framework de traitement du langage utilisable notamment pour classifier des documents et labelliser

[1680] Je décris leur approche dans l'ebook Comprendre l'informatique quantique, septembre 2019 (504 pages).

des séquences de vidéos, **Ax** (Adaptive Experimentation Platform), une plateforme de configuration et de paramétrage de tests lors des développements et **BoTorch**, une bibliothèque de gestion de probabilités Bayésiennes. De nombreux autres outils open source ont été publiés par Facebook, Glow (compilateur de deep learning), FAISS (recherche de contenus multimédias similaires), StarSpace (classification de textes), Visdom (dataviz), DynaBench (analyse dynamique de jeux de données), Translate (traduction) et VizSeq (génération de textes).

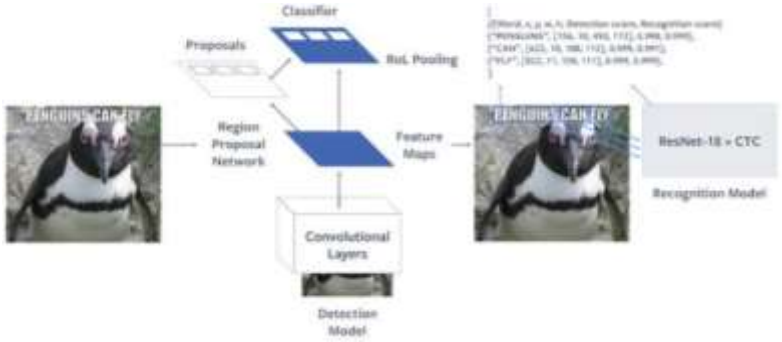
La société a plus de 600 chercheurs en IA[1681] dont le fameux Yann Le Cun, créateur des premiers réseaux de neurones convolutifs qui sont la base de la reconnaissance d'images dans le deep learning.

Il gère la Direction Scientifique de l'équipe de chercheurs du FAIR (Facebook Artifical Intelligence Research) dont le patron est un autre français, Jérôme Pesenti, qui reporte au CTO de Facebook. Les laboratoires de FAIR sont installés à Menlo Park en Californie, New York, Paris (dirigé par Antoine Bordes[1682]), Montréal (dirigé par Joëlle Pineau), Londres, Tel Aviv, Seattle et Pittsburgh. Les trois premiers laboratoires emploient environ 60 chercheurs et celui de Montréal une vingtaine. Mais ce ne sont pas les seuls laboratoires de recherche de Facebook.

Parmi les projets de recherche de Facebook, on compte **DeepFace**, une solution de reconnaissance des visages avec un réseau de neurones de neufs couches et 120 millions de connexions, entraîné sur quatre millions d'images. La précision du système serait de 97%. Le système a bénéficié de la contribution de Yaniv Taigman, issu de **Face.com** (2005, Israël, $11M), acquise en 2007.

Facebook veut notamment utiliser la reconnaissance d'images pour informer ses utilisateurs malvoyants du contenu visuel de leur timeline.

D'autres projets de recherche marquants comprennent **Rosetta** qui reconnaît les textes dans les images et vidéos[1683] et des outils pour aider les développeurs à détecter **(SapFix)** et corriger **(Sapienz)** les bugs dans leur code[1684]. Reste à en apprécier la portée exacte !



Les autres acquisitions de Facebook dans l'IA comprennent :

- **Jibbigo** (2009, USA) en 2013, pour son système de traduction speech-to-speech.
- **Pebbles Interfaces** (2010, Israël, $14,45M) en 2015, pour son système de captation de gestes.

[1681] La liste des chercheurs de Facebook est publiée ici : https://research.fb.com/people/page/33/?paginated=true. Données du 31 août 2019.

[1682] Le laboratoire FAIR de Paris comprend plus de 80 collaborateurs dont une trentaine de doctorants Cifre. Il a aussi cofinancé en 2019 le supercalculateur Jean Zay du GENCI à hauteur de 3M€, couvrant l'acquisition de plusieurs centaines de GPGPU Nvidia. Voir Cinq ans après son lancement, Facebook fait le bilan de son laboratoire d'intelligence artificielle à Paris par Julien Bergounhoux, juin 2020. Voir aussi À Paris, Facebook compte créer une IA douée de raison par Antoine Crochet-Damais, juin 2020, qui évoque les projets menés par le FAIR Paris.

[1683] Voir Rosetta: Understanding text in images and videos with machine learning, septembre 2018.

[1684] Voir Finding and fixing software bugs automatically with SapFix and Sapienz, septembre 2018.

- **wit.ai** (2013, USA, $3M) en 2015, une petite startup de Palo Alto, pour ajouter des fonctionnalités de reconnaissance de la parole dans ses services et notamment Messenger. Mais Wit.ai est aussi une plateforme utilisée par des milliers de développeurs. La startup avait été créée par trois français : Alexandre Lebrun, Willy Blandin et Laurent Landowski.
- **Surreal Vision** (2014, UK), une startup acquise par Oculus en 2015, spécialisée dans vision 3D.
- **Faciometrics** (2015, USA) en 2016, pour sa solution mobile d'analyse de visages.
- **Ozlo** (2014, USA, $14M) fin juillet 2017, qui aide à trouver un bon restaurant.
- **Blomsbury** (2015, USA, $1,7M) en juillet 2018, qui permet d'interroger en langage naturel des documents non structurés. C'était l'objet de leur projet Cape, publié en open source fin 2017. L'outil doit en théorie permettre de lutter contre la propagation de fake news.
- **GrokStyle** (2016, USA, $2,5M) en février 2019, qui sert à reconnaître des objets dans un catalogue produit avec son smartphone. Leur outil était intégré dans l'application mobile d'Ikea.

Alors qu'il se refusait à censurer les publicités politiques remplies de mensonges, Facebook lancait en septembre 2019 le **Deepfake Detection Challenge**, doté de $10M de prix. Il s'agit de créer un système de détection d'images générées par des IA[1685].

Les données

Facebook est un gros collecteur de données générées par les utilisateurs de ses services.

- Le **texte** via Facebook et Whatsapp.
- Les **photos** via Facebook et Instagram.
- Les **vidéos** via Facebook.

En règle générale, Facebook ne vend pas de données mais l'accès à des utilisateurs en fonction de profiling issus de ces données. Mais l'affaire Cambridge Analytica au printemps 2018 a montré que le système avait des trous dans la passoire, les éditeurs d'applications pouvant dans certaines conditions récupérer des informations de profiling des utilisateurs. Depuis, Facebook a fait le ménage dans ses APIs mais rien n'est sûr à 100%.

Facebook entraîne ses systèmes de reconnaissance d'image avec 3,5 milliards de photos !

Le matériel

En octobre 2018, Facebook faisait sa dernière incartade dans le matériel[1686] avec le lancement de **Portal**, un écran pour les communications vidéo supportant aussi Amazon Alexa.

Un cache est fourni pour sa caméra pour préserver la vie privée. Facebook insiste beaucoup sur le « privacy by design » intégré dans l'objet pour protéger la vie privée des utilisateurs.

Ce qui n'a pas empêché les échos négatifs dans une partie des médias, surtout aux USA[1687].

[1685] Voir Creating a data set and a challenge for deepfakes par Mike Schroepfer, septembre 2019.

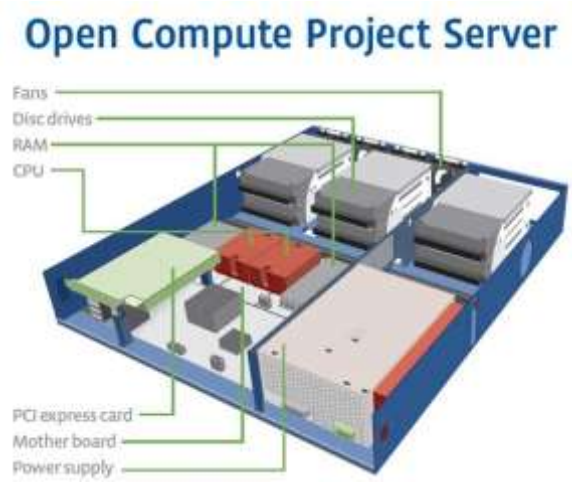
[1686] Après, bien entendu, son acquisition d'**Oculus Rift** (2012, USA, $96M) en 2014.

[1687] Voir Facebook's launch of Portal has been stymied by trust issues de Casey Newton, octobre 2018.

Comme tous les grands acteurs de l'Internet, Facebook conçoit ses propres serveurs pour ses data-centers et en particulier pour les applications exploitant de l'IA qui sont gourmandes en puissance de calcul. Ils ne vont cependant pas jusqu'à concevoir leurs propres processeurs ASIC ou FPGA comme le font Google, IBM et Microsoft. Facebook utilise des processeurs standards du marché provenant surtout d'Intel et de Nvidia[1688].

Ils ont donc conçu leurs propres *server appliances* dédiés à différentes tâches en séparant notamment les fonctions d'entraînement et d'inférence. Leurs serveurs d'entraînement **Big Basin** étaient mis en production en 2017 [1689] (*ci-contre*). Equipés de 8 GPU Nvidia V100, ils sont équivalents, en plus denses, aux serveurs Nvidia DGX-1.

Leur infrastructure a depuis bien évolué[1690]. Facebook a contribué avec Microsoft à la création de l'**Open Compute Project**, une architecture ouverte de serveurs pour datacenters qui définit les grands composants de serveurs de data-centers.



Leurs serveurs d'entraînement adoptent l'architecture **Zion**, adaptable à tous types de réseaux de neurones et des modules d'accélération OCP (OAM pour OpenCompute Accelerator Modules) exploitant divers processeurs du marché. En plus de ceux d'Intel et Nvidia, ils utilisent aussi ceux d'**AMD**, **Habana** et **Graphcore**. Ces serveurs comprennent chacun huit modules OAM (*ci-dessous*).

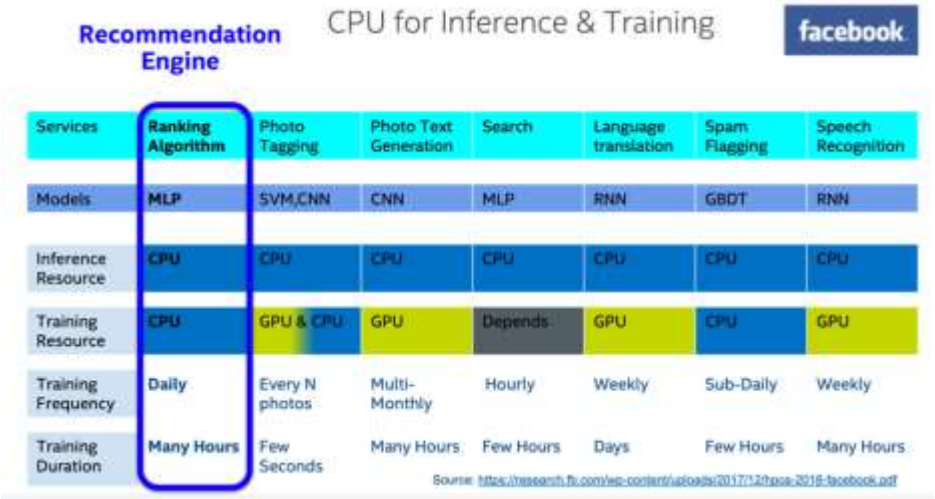
[1688] Voir ML at Facebook: An Infrastructure View par Yangqing Jia, 2018 (38 slides) et How Facebook Scales Machine Learning par Jamal Robinson, février 2019.

[1689] Voir Introducing Big Basin: Our next-generation AI hardware, mars 2017.

[1690] Voir Accelerating Facebook's infrastructure with application-specific hardware, mars 2019.

Les serveurs d'inférences dénommés **King Canyon** utilisent des processeurs qui fonctionnent en calculs entiers (INT8, sur 8 bits) ou flottants (FP16), selon les besoins.

L'architecture modulaire s'appuie sur des modules d'inférence avec connecteurs M.2 (les mêmes que ceux des SSD de vos laptops), des modules les intégrant (Glacier Point), des cartes serveurs classiques (Twin Lakes avec processeurs Intel), le tout dans un châssis (Yosemite)[1691].



Facebook utilise aussi le machine learning dans la gestion de ses data centers[1692].

Apple

Apple est bien plus orienté produits et marchés que ne le sont IBM et Microsoft. Non seulement la société n'a pas formellement de laboratoire de recherche fondamentale mais elle ne publiait jusqu'en 2016 *aucun* papier dans le domaine de l'IA. C'est tout le contraire de l'innovation ouverte !

Elle est cependant sortie du bois en publiant un premier papier fin 2016 sur la reconnaissance d'images[1693]. La suite se trouve dans leur Apple Machine Learning Journal.

L'IA est maintenant incontournable dans l'offre d'Apple, tant au niveau logiciel (SIRI, photos) que matériel (chipsets A11, A12 et A13 avec NPU, HomePod, incartades dans les véhicules autonomes). Son influence reste cependant plutôt marginale dans l'IA face aux autres GAFAMI, en particulier auprès des développeurs d'applications d'entreprises.

[1691] Source du schéma : The Cambrian AI Landscape: Intel par Karl Freund dans Forbes, février 2021.

[1692] Voir Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective, 2017 (11 pages).

[1693] Voir Learning from Simulated and Unsupervised Images through Adversarial Training en décembre 2016.

Les logiciels

Les acquisitions d'Apple sont peu nombreuses en règle générale même si elles ont eu tendance à s'accélérer depuis le décès de Steve Jobs en 2011.

Dans l'IA, on peut compter **Turi** (2013, USA, $25M) en 2016 pour $200M qui avait développé un projet d'analytics à base de machine learning, **Emotient** (2006, USA, $6M) en 2016 pour la reconnaissance des visages et des émotions, **VocalIQ** (2011, UK, $1,2M) en 2015 qui devait enrichir les fonctionnalités de reconnaissance de la parole de SIRI en ajoutant de l'auto-apprentissage, ainsi que **Perceptio** (2014, USA) en 2015, dans la reconnaissance d'images s'appuyant sur du deep learning et **Regaind** (2015, France, 400K€) en septembre 2017.

Ont suivi les acquisitions de **Lattice Data** (2015, USA) en 2017, issu du projet DeepDive de Stanford, spécialisé dans l'extraction de données de contenus non structurés, et **SensoMotoric Instruments** (1991, Allemagne) qui a développé une technologie de suivi du regard à base d'IA (eye tracking).

SIRI est de son côté le résultat de l'acquisition en 2010 de la startup **SIRI** (2007, USA, $24M) en 2010, elle-même issue d'un projet de **SRI International** financé par la DARPA, et de l'usage des technologies issues de l'américain **Nuance Communications**, la société leader du secteur de la reconnaissance de la parole qui fait plus de $2B de chiffre d'affaires !

Ce dernier utilise en partie des technologies issues de Scansoft, provenant du belge Lernout & Hauspie qui avait acquis la technologie de reconnaissance de la parole de Ray Kurzweil !

Apple comble les trous dans son offre d'IA via son partenariat avec IBM qui porte notamment sur Watson, une manière indirecte de séduire les grandes entreprises et les DSI, qui étaient les grandes bêtes noires de Steve Jobs.

Il est cependant probable qu'Apple devra faire quelques acquisitions dans le cadre de son projet de voiture autonome.

Apple lançait son framework **CoreML** sur iOS en juin 2017. C'est un framework qui permet l'exécution de modèles d'IA déjà entraînés sur smartphones. La version 2 de Core ML a été lancée un an plus tard. C'est un framework généraliste qui permet de gérer des modèles entraînés de deep learning aussi bien que de machine learning classique avec des régressions linéaires, non linéaires, du SVM et autres modèles probabilistes du genre.

Apple faisait l'acquisition de **Silk labs** en 2019 (2015, USA, $2,5M). Cette startup leur apporte des briques logicielles permettant d'exécuter des solutions de deep learning dans des objets connectés avec une faible empreinte mémoire et processeur.

Les données

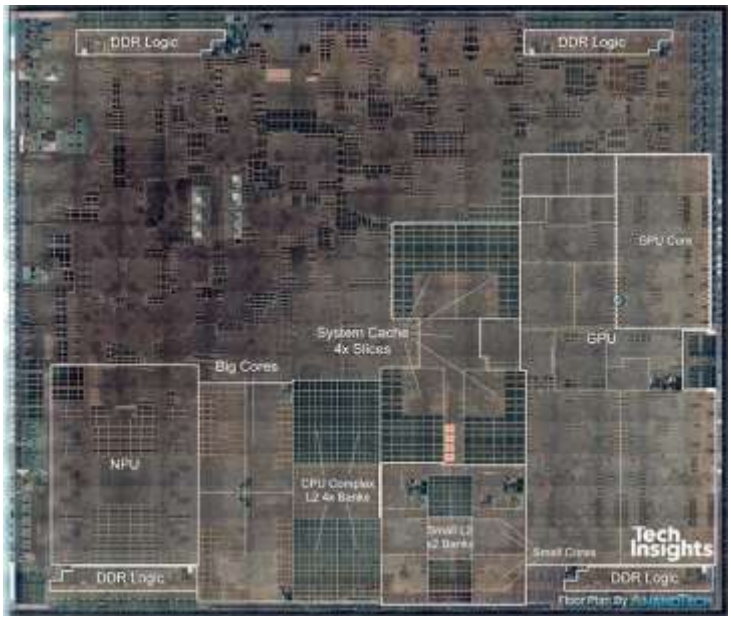
Apple gère de gros volumes de données de ses utilisateurs, que cela soit via iCloud ou son magasin d'applications mobiles qui lui permettent de connaître pas mal de choses sur leurs habitudes. Ceci étant, ces données ne sont pas utilisées par Apple avec la même logique qu'un Google ou un Facebook. Apple se positionne même explicitement comme étant plus respectueux de la vie privée de ses utilisateurs.

Derrière les bonnes intentions, l'origine en est limpide : c'est lié à leur modèle économique de vente de matériel qui domine leur mix de revenu alors que Google et Facebook dépendent presque exclusivement du revenu publicitaire.

Le matériel

Apple utilise beaucoup d'IA dans ses iPhone. Les iPhone 8 et X annoncés en septembre 2017 intègrent le chipset A11 Bionic et sa fonctionnalité neuromorphique Neural Engine, dédié à l'exécution d'applications de deep learning, comme pour la fonction FaceID de reconnaissance de visage ainsi que pour la reconnaissance de la parole avec SIRI.

La série d'iPhones Xs lancée en septembre 2018 intègre le chipset A12 réalisé en technologie 7 nm avec un doublement des capacités de son NPU (*ci-contre*). Le NPU est le rectangle en bas à gauche du processeur[1694]. Il a une capacité d'environ 4 Tops (téra opérations par seconde, sur des entiers). Il est principalement exploité par les fonctions photos, de login par la détection du visage et de reconnaissance vocale des smartphones. Les iPhones 11 de septembre 2019 sont équipés des chipsets A13 qui passent à 8 cœurs, dont la fonction n'est pas précisée plus des multiplicateurs de matrices.



Suivaient les chipsets A14 des iPhone 12 de 2020 avec une puissance accrue de leur NPU, atteignant 11 TOPS avec 812 millions de transistors. Sans compter le processeur M1 destiné aux Macbooks lancés fin 2020, qui est lui aussi équipé d'un Neural Engine à 16 cœurs de nature non précisée et totalisant 11 TOPS, exactement comme dans le A14.

Apple a sinon lancé ses enceintes à commande vocale équipées de SIRI, les Homepod en 2018. Etant dans le troisième écosystème du marché des enceintes vocales après ceux d'Amazon Alexa et de Google Assistant, leur marché est surtout celui des Apple aficionados.

Enfin, Apple investi dans le domaine des véhicules autonomes. On a commencé à en entendre parler presque officiellement avec le projet Titan, qui connaît cependant des hauts et des bas. Il semblerait que leur approche dans un premier temps soit celle de l'équipementier, créant des systèmes de conduite autonomes avec matériel et logiciel, destinés à des constructeurs automobiles. Cela fait longtemps que des rumeurs circulent sur leurs velléités de construire des véhicules, ou de faire l'acquisition de Tesla. Mais elles n'ont pas été confirmées à ce jour.

Salesforce

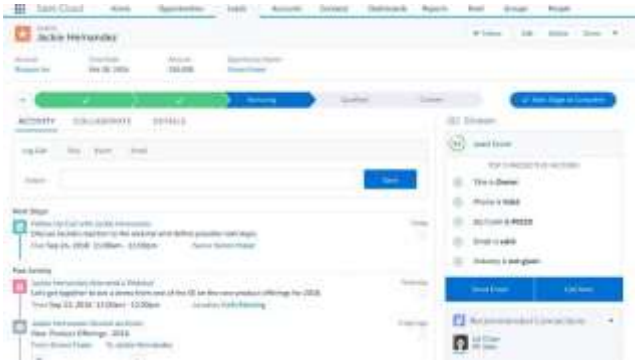
Chez Salesforce, l'offre d'IA s'appelle modestement **Einstein**, une offre d'IA en cloud au service des forces de vente.

L'offre qui s'appelle précisément **Einstein High Velocity Sales Cloud** comprend notamment les briques suivantes, la liste s'allongeant régulièrement[1695] :

[1694] Source de l'image : The iPhone XS & XS Max Review: Unveiling the Silicon Secrets de Andrei Frumusanu, octobre 2018.

[1695] Voir Salesforce ajoute une autre dose d'IA à Sales Cloud par Maryse Gros, avril 2019.

- **Einstein Lead Scoring** : avec des outils à base de machine et deep learning de repérage des meilleurs leads d'un pipe commercial en fonction d'une analyse multicritères (pour peu que la base soit bien renseignée…). Un cas d'usage classique du machine learning en tout cas.
- **Einstein Activity Capture** : capture des informations utiles dans les mails et calendriers et des modèles personnalisés de réponses par e-mails.
- **Lightning Sales Console** : un espace de travail personnalisable de suivi des meilleurs leads.
- **Lightning Dialer** : pour contacter les prospects en un clic.
- **Lightning Web Components** est une boîte à outils pour développeurs qui comprend maintenant une fonction d'OCR[1696].
- **Salesforce Engage** : notifications en temps réel d'opportunité d'interaction client.
- **Salesforce AppExchange** : un écosystème de solutions tierces-parties.



- **Einstein Bots**, sa plateforme de création de chatbots lancée en novembre 2017 qui exploite **Einstein Language**, l'outil de traitement du langage de Salesforce. Le système détecte le niveau d'émotion de l'utilisateur (positif, négatif ou neutre). Einstein Bots comprend une interface graphique pour la conception de chabots. Ces chatbots sont destinés aux clients de l'entreprise exploitant Salesforce tout comme à leurs commerciaux.
- **Einstein Visual Search**, une fonction de reconnaissance visuelle de produits qui est intégrable dans des catalogues de ventes en ligne et fait partie de l'offre **Commerce Cloud**.
- Des fonctions de **traduction** utilisables notamment pour préparer le travail de centres d'appels internationaux.
- **Einstein Prediction Builder**, un système de création de prévisions applicable à toute donnée gérée par Salesforce. Et pour ceux qui veulent développer, Salesforce faisait un don à la communauté open source qui demande le respect avec **AutoML TransmogrifAI** qui comprend 22 lignes de code, commentaires compris. Elle s'appuie sur Apache Spark et est écrite en Scala[1697].

La recherche en IA chez Salesforce s'appuie sur au moins 100 ingénieurs pour la recherche fondamentale et plus de 200 ingénieurs pour la recherche appliquée.

Une bonne partie d'Einstein provient de briques et de compétences récupérées par SalesForce à l'occasion de l'acquisition de diverses startups : **RelateIQ** (2011, USA, $69M) en 2014, qui était spécialisées dans la relation client, devenu SalesForceIQ, une version de SalesForce pour les PME, **PredictionIO** (2013, USA, $2,7M) acquis en 2016 pour sa solution open source de machine learning et **MetaMind** (2014, USA, $8M) acquis en 2016 pour ses solutions de reconnaissance d'image, soit l'équivalent de 175 data scientists. Einstein s'appuie aussi sur les APIs d'**IBM Watson**.

[1696] Voir Salesforce adds OCR, translation, AI app builder to Einstein platform, avril 2019.

[1697] Ce n'est pas une blague. Le code est ici : https://transmogrif.ai/.

Enfin, en 2019, Saleforce faisait l'acquisition de **Tableau** pour $15,7B, ajoutant à son arc des analytics et une bonne dose de machine learning pour analyser la structure des données en septembre 2019. L'outil proposé a l'air de faire de l'analyse de composantes (PCA) pour identifier les corrélations entre données source et données cibles d'un tableau structuré.

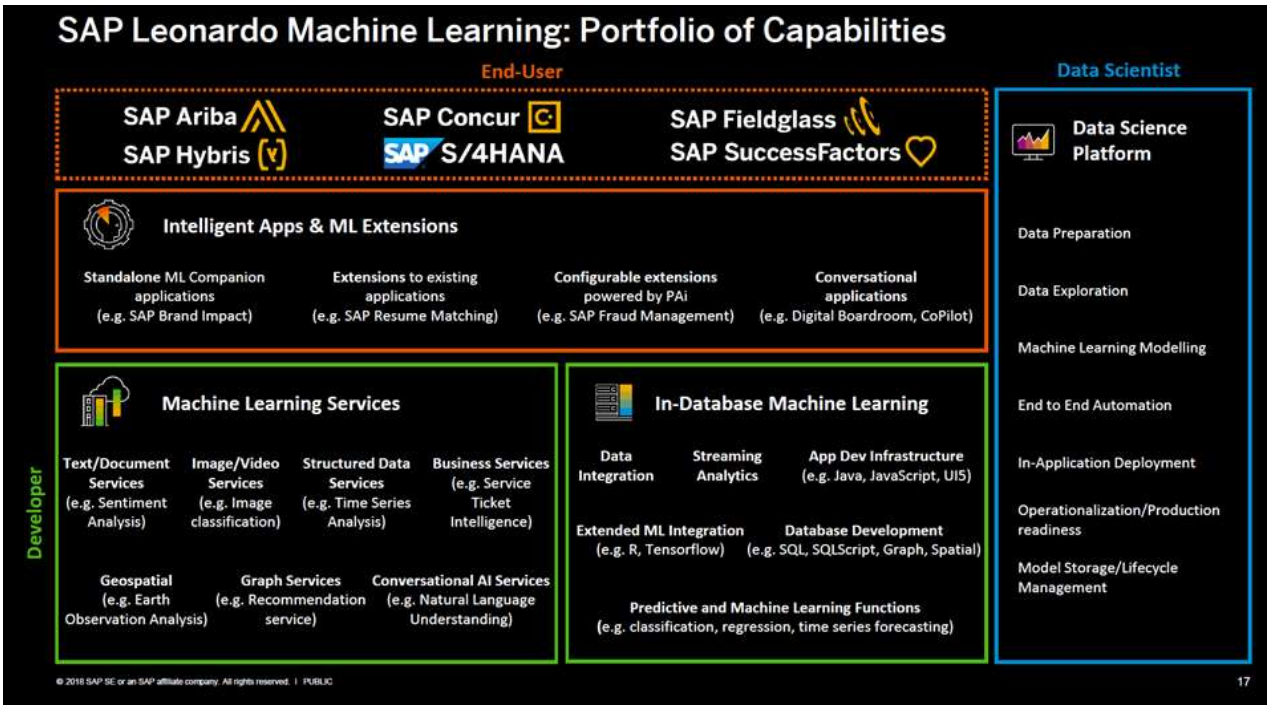
SAP

L'IA de SAP est intégrée depuis l'été 2017 dans le vaste gloubi-boulga qu'est la plateforme **Leonardo**. Les équipes marketing de SAP méritent largement d'aller dans l'enfer des marketeurs tellement les évolutions de leur offre en IA sont difficiles à suivre avec des dizaines de briques difficiles à positionner les unes par rapport aux autres.

Les briques logicielles de Leonardo comprennent l'ensemble des outils servant à exploiter les données générées par les applications métiers transactionnelles de SAP. Leonardo s'appuie sur divers outils de machine learning autour de **SAP Leonardo Machine Learning Foundation** (vidéo) qui couvre tous les besoins de gestion de données, de textes et d'images. La plateforme comprend aussi un moteur de gestion de règles (Rules Engine). Cette plateforme de machine learning s'appelait précédemment Clea.

Côté outils de travail, on y trouve notamment l'APL (**Automated Predictive Library**) pour les analystes et la PAL (**Predictive Analysis Library**) qui est destinée aux data scientists.

Côté couches basses, Leonardo supporte aussi l'intégration de modules développés en **TensorFlow**. SAP fait la promotion du « *in-database machine learning* », même si ce niveau d'intégration n'est pas bien précisé ni à quelle base de données cela correspond. Cela semble en fait relever du « *in-memory processing* ».



Il s'agit en tout cas de machine learning s'appuyant sur des méthodes classiques : arbres de décision, SVM, K-means, régressions linéaires et non linéaires, etc. Les modèles plus complexes avec des réseaux de neurones et du deep learning sont mis en œuvre avec Tensorflow.

SAP décline toutes ces plateformes dans de nombreux cas d'usage et marché verticaux et en particulier dans l'industrie et les objets connectés.

L'éditeur exploite même des outils de reconnaissance d'images pour certains besoins, tels **SAP Brand Impact** qui mesure la visibilité de marques dans les médias ou dans un stade ou **SAP Resume Matching** pour la recherche de correspondance de CV et de postes ouverts dans la RH.

En 2019, SAP faisait l'acquisition de **Contextor** (2000, France, 600K€), ajoutant à son offre une solution de RPA. Cela complète dans un autre domaine l'acquisition de **Recast.ai** (2015, France, 1M€) en 2018 avec son outil de création d'agents conversationnels.

Enfin, côté serveurs, SAP est partenaire à la fois d'**Intel** et **Nvidia** pour l'exécution des briques de machine learning de Leonardo, en particulier sur les serveurs DGX de Nvidia.

Oracle

Avec SAP, Oracle est l'un des plus gros éditeurs de logiciels d'entreprise mondiaux. Il était plutôt discret dans son intégration de l'IA dans la stratégie jusqu'en 2017. Ils se sont embarqués dans l'IA en intégrant progressivement et ouvertement divers outils de machine learning dans leurs logiciels d'infrastructures horizontaux et applications métiers verticales[1698]. A commencer par la quincaillerie permettant de créer des chatbots, un type de projet d'IA très courant dans les entreprises.

On en trouve ainsi dans **Oracle Management Cloud Services**, dans **Oracle Advanced Analytics, Oracle Data Miner** et **Oracle Internet of Things Cloud Service**, qui sont des outils d'analytics divers, exploitant des arbres de décision et générant des rapports divers.

Oracle annonçait son **Chatbots for Developers** en 2016 et une plateforme de développement associée et le service associé, **Intelligent Bot Cloud Service** (IBCS). Leur outil permet de créer des chatbot à commande textuelle ou vocale. Je n'ai pas trouvé son origine. Il fait évidemment appel aux outils de gestion de bases de données et de cloud d'Oracle. En octobre 2018, Oracle mettait à jour son offre de chatbot avec une solution d'assistant numérique intégrant l'accès aux applications d'ERP, de CRM ainsi que de RH.

Cet Oracle Digital Assistant est commandable par la voix (via Amazon Alexa) ou par le texte (via Slack, Facebook Messenger, WeChat) (vidéo).

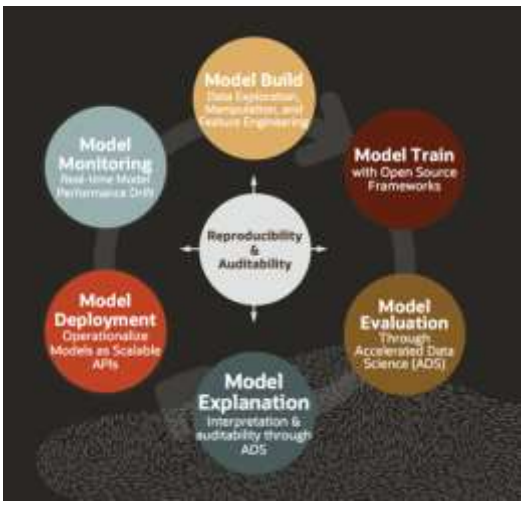
Début 2019, Oracle saupoudrait d'IA ses différentes solutions applicatives et métiers. Un meilleur chatbot pour faire ses notes de frais dans son ERP, une automatisation des audits financiers à base de machine learning, le scoring de fournisseurs dans leur offre supply chain issu de l'acquisition de la startup **Datafox** (2013, USA, $13,6M) en novembre 2018, du machine learning intégré dans la planification de projets, du machine learning dans Oracle SCM, leur solution de supply chain, des assistants texte et vocaux ainsi qu'un outil de planification de ressources humaines à base de machine learning pour Oracle HCM, leur solution RH, une planification améliorée du cycle de vente dans leur offre de CRM qui rappelle ce que fait Salesforce avec Einstein.

En 2019, Oracle lançait aussi **Exadata Database Machine X8**, un serveur de base de données « rack » pour data center complet comprenant des fonctionnalités de machine learning, notamment autour de l'indexation, afin d'optimiser l'entraînement des modèles en fonction de l'évolution des données utilisées. Chaque rack intègre jusqu'à 912 cœurs de CPU avec jusqu'à 3 Po de stockage et 28 To de mémoire vive pour gérer les bases SQL en RAM.

[1698] Accenture et Oracle publiaient un livre blanc décrivant la stratégie IA d'Oracle, Technology Vision for Orange Executive Summary (32 pages). Ca parle d'entreprise intelligence et d'IA humaine. Mais franchement, pas beaucoup d'IA !

Quid des acquisitions dans l'IA ? Oracle est assez friand de startups, mais curieusement, pas vraiment dans l'IA. On note surtout celle de **Crosswire** (2013, Israël, $5M) en 2016 qui propose une solution cross-device de ciblage et d'analytics publicitaires doté d'outils de présentation graphique. Et puis celle de **Datascience** (2014, USA, $28M) en juin 2018, un spécialiste de la data science et du machine learning.

Oracle lançait en février 2020 son nouveau service **Cloud Infrastructure Data Science**, reposant sur l'Oracle Cloud Infrastructure, destiné aux data scientists pour leur permettre de collaborer autour de projets de machine learning allant du développement, au déploiement et à la maintenance.



L'outil permet d'accéder aux briques d'autoML maison d'Oracle ainsi que d'autres servant à l'explication des modèles de machine learning générés via la bonne traçabilité des données d'entraînement. L'environnement est accédé au travers de Jupyter et est compatible avec les principaux frameworks du marché tels que TensorFlow, Keras et SciKit-Learn.

Cisco

Le grand spécialiste des réseaux d'entreprises s'est évidemment aussi mis à l'IA. Nous les avons déjà cités dans cet ebook : ils ont une offre de serveurs intégrant des chipsets Intel et Nvidia V100 pour les data-centers ainsi que des solutions de cybersécurité s'appuyant sur des briques d'IA.

Ils exploitent sinon le machine learning dans la plupart des grandes briques logicielles de leur offre[1699]. Leur solution ETA (Encrypted Traffic Analytics) détecte les menaces cyber en analysant le trafic réseau encrypté grâce au machine learning.

Principalement, en exploitant les métadonnées de ce trafic (origine, fréquence, forme des trames, …).

Startups

L'IA est devenue un phénomène de mode pour les startups comme l'ont été les réseaux sociaux (vers 2004), la vidéo en ligne (vers 2006), la mobilité (à partir de 2009), les objets connectés (vers 2011), le cloud ou la Blockchain (depuis 2017).

Depuis 2015, une startup qui ne brandit pas l'IA comme sauce magique paraît dépassée par les événements. Et les sous-modes abondent avec les chatbots, la robotique, le cognitif, etc.

[1699] Source de l'illustration: Cisco has the edge with artificial intelligence and machine learning par Stephanie Chan, juillet 2018.

Caractéristiques

On sourit souvent des startups qui ont "créé une IA" alors qu'elles ont correctement paramétré un réseau de neurones avec TensorFlow à partir d'exemples et après avoir tâtonné, ou qu'elles ont utilisé une vieille technique de prévision ou de clustering à base de machine learning.

Ces effets de mode sont notamment alimentés par les prévisions de chiffre d'affaires des analystes, comme IDC qui prédit que le marché de l'IA représentera $46B de CA en 2020[1700]. Ces évaluations ont toujours tendance à gonfler des chiffres qui sont incalculables. Ainsi, quel est le CA en IA de Google, Facebook ou Microsoft, qui ne sont d'ailleurs visiblement pas intégrés dans l'estimation d'IDC ? Une donnée intéressante serait d'évaluer le CA additionnel généré par l'IA chez les éditeurs de logiciels et startups, mais la part de l'IA dans la valeur ajoutée d'un Oracle ou d'un SalesForce est bien difficile à évaluer. On a connu le même phénomène avec les prévisions sur le marché des objets connectés[1701].

Cela amène la généralisation du phénomène de l'*IA washing*, décrivant ces startups qui usent et abusent de la terminologie de l'IA, souvent, sans préciser la manière dont leur solution en tire parti. Ce qui ne veut pas dire qu'elles ne font pas d'IA pour autant. Elles le font à des degrés très variables. Les techniques de l'IA sont largement disponibles, en open source, dans le cloud et dans l'embarqué. Les techniques du machine learning sont relativement faciles à mettre en œuvre. Celles du deep learning nécessitent un effort conceptuel plus important, mais accessible aux développeurs et data scientists sans compter les boîtes à outils prêtes à l'emploi comme chez Clarifai pour l'analyse d'images. L'IA est en train de devenir l'équivalent moderne du développement web : un ensemble de techniques de plus en plus abordables.

Si une entreprise indique avoir « *créé une IA* », il faut traduire par « *nous avons créé une solution logicielle et/ou matérielle qui intègre des briques technologiques de l'IA et souvent d'autres briques techniques* ». Pour d'autres, une IA est un système anthropomorphique qui imite une caractéristique humaine, comme la compréhension du langage. Pour moi, c'est un logiciel qui intègre une ou plusieurs briques logicielles du vaste champ scientifique de l'IA depuis sa création en 1955. Cela intègre le machine learning, le deep learning, les moteurs de règles tout comme les réseaux multi-agents. Et une IA logicielle n'est pas forcément anthropomorphique. Elle peut réaliser des opérations qui sont inaccessibles à l'intelligence et à la mémoire humaines. Un système de segmentation automatique d'une énorme base de données réalise une tâche surhumaine !

C'en est au point où comme l'Internet ou la mobilité, l'IA est en train de devenir incontournable. Progressivement, l'ensemble des solutions logicielles du marché vont intégrer des briques d'IA. N'importe quel logiciel de gestion ou logiciel métier, par exemple, utilisera le machine learning pour présenter des analyses des données générées et faire des prévisions. Tout logiciel exploitant du texte intégrera des fonctions de traitement du langage. Tout logiciel exploitant des images va s'appuyer sur du deep learning pour déterminer ce que contiennent les images.

En mars 2019, le fonds d'investissement britannique **MMC Venture** défrayait la chronique en publiant une étude de 151 pages dont quelques lignes faisant état du fait que 40% des startups européennes prétendant faire de l'IA n'en faisaient pas.

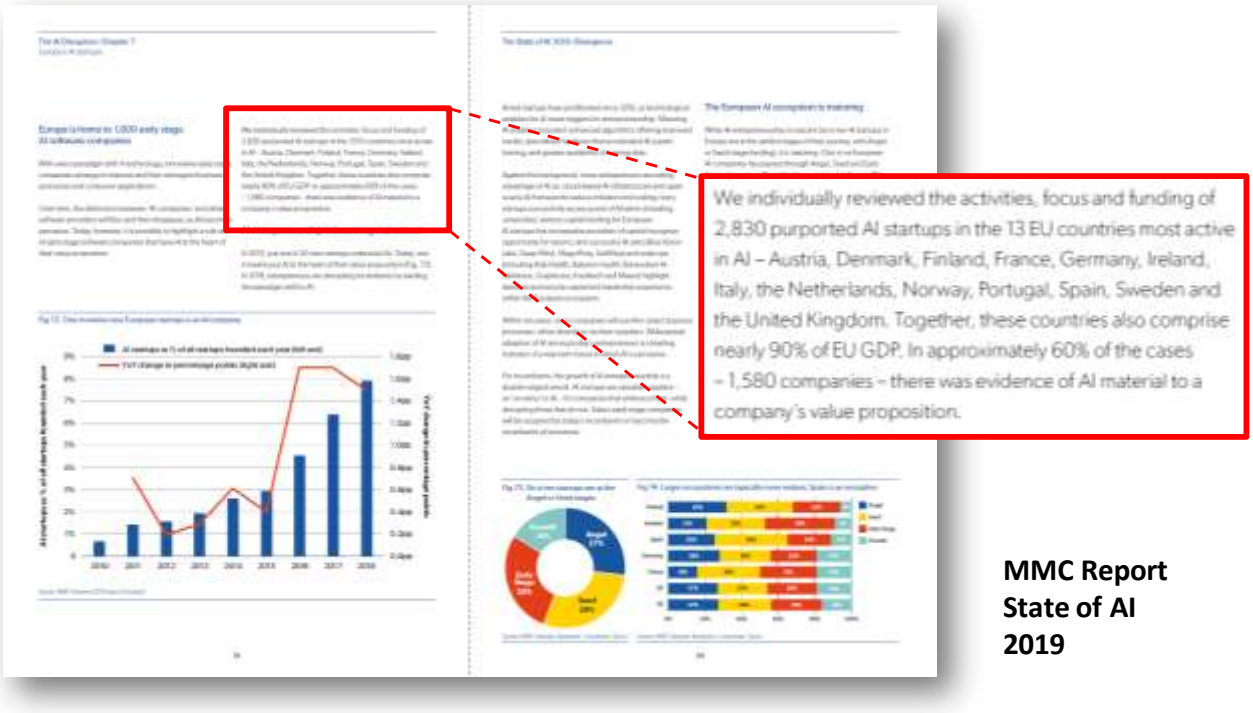
L'étude était fortement biaisée avec une définition très restrictive de l'IA qui excluait d'emblée tout ce qui ne relevait pas du machine learning si ce n'est du deep learning et un jeu de données de startups non public[1702].

[1700] Voir Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $12.5 Billion This Year, According to New IDC Spending Guide, avril 2017.

[1701] Que j'avais eu l'occasion de décrire dans La grande intox des objets connectés en août 2015.

[1702] J'ai décrit l'ensemble des biais de cette étude dans Les startups font-elles de la fausse IA ?, Olivier Ezratty, mars 2019. Voir aussi AI: Fake it 'till you make it par Pierre-Julien Grizel, 2019, qui décrit bien les nuances à adopter dans l'évaluation d'une startup faisant de l'IA.

L'étude servait surtout à valoriser quelques participations de ce fonds, presque exclusivement issues du Royaume-Uni.



Le tout, juste avant qu'il lance un nouveau fonds de £52M quelques semaines plus tard[1703]. Tout ceci doit inciter à la prudence dans la reprise de chiffres ainsi publiés.

Laurent Alexandre aime à dire que les investissements dans les startups de l'IA ne valent pas grand-chose et que leurs investisseurs sont des gogos, les seuls créateurs de véritables IA étant les GAFA. C'est évidemment exagéré ! Nombre de startups créent des IA aussi sophistiquées que celles des GAFA. En fait, le risque est simplement que l'usage de l'IA dans les logiciels devienne une simple commodité, donc une banalité. Comme le fait de créer un logiciel sous la forme d'un couple entre serveur web et application exploitable à partir d'un navigateur. D'ici 5 à 10 ans, le marché de l'IA se confondra donc quasiment avec l'ensemble du marché du logiciel et d'Internet.

Les GAFA créent et utilisent des IA sous la forme de logiciels open source et de data centers. Les algorithmes innovants sont encore créés en masse par des chercheurs issus d'universités du monde entier. Les GAFA n'ont pas encore le monopole de la créativité dans la recherche. Qui plus est, les solutions techniques sont open source et ne doivent pas obligatoirement tourner dans les infrastructures des GAFA. Comme l'Internet, l'IA est tout de même très distribuée.

Derrière l'habillage marketing, il reste à comprendre ce que la startup a réellement produit : a-t-elle assemblé des briques logicielles existantes de manière traditionnelle, a-t-elle créé des briques spécifiques, a-t-elle juste entraîné un modèle assez simple et mis en forme des données, d'où viennent-elles et la solution est-elle une simple application directe de techniques existantes ? En général, c'est bien le cas. Mais le choix, la programmation et l'entraînement d'un modèle de deep learning ou de machine learning pour répondre à un besoin spécifique requiert des compétences encore rares, sans compter les données d'entraînement.

On ne devient pas développeur dans l'IA du jour au lendemain, de même qu'il a fallu du temps pour que les développeurs d'applications procédurales ou client-serveur s'adaptent à la programmation événementielle du web et avec ses nombreux framework qui changent tout le temps (jquery, Angular, React, Node).

[1703] Voir MMC Ventures launches fresh £52M seed fund aimed at London startups par Mike Butcher, mai 2019 et l'équivalent en France : 40% des start-up européennes d'intelligence artificielle n'utilisent pas d'intelligence artificielle par Elisa Braun, mars 2019.

Selon IDC, 1% des logiciels utiliseraient de l'IA aujourd'hui et en 2018, 75% des développeurs intégreront de l'IA dans leur code, ce qui est probablement un peu optimiste, ne serait-ce que pour tenir compte du laps de temps pour se former[1704].

L'IA en est encore au stade artisanal et du bricolage. Cela ne se voit évidemment pas directement quand on fait le tour d'horizon des startups du secteur. Surtout dans la mesure où la plupart d'entre elles sont "b-to-b" et diffusent leurs solution en marque blanche. Vous les retrouverez éventuellement dans les agents conversationnels des sites web de marques, dans le ciblage marketing qui vous touche avec une offre pertinente (ou pas du tout…), dans des robots capables de dialoguer plus ou moins avec vous, ou dans les aides à la conduite de votre voiture haut de gamme.

L'un des moyens de se rendre compte indirectement de cet aspect artisanal consiste à évaluer la part produit et la part service des entreprises du secteur. Plus la part du produit est faible, plus on est dans le domaine de l'artisanat. Cela n'apparaît pas dans les données publiques mais peut au moins d'obtenir quand on a l'occasion d'observer à la loupe ces entreprises : dans le cadre d'une relation grand compte/startup, d'un investissement ou même d'un recrutement. On peut l'observer également dans les profils LinkedIn des salariés de l'entreprise s'ils sont disponibles. Bref en utilisant ce que l'on appelle des sources d'information "ouvertes".

Les startups de l'IA, surtout américaines, ont quelques points communs marquants :

- Elles ont majoritairement des **approches marché "b-to-b"** avec des marchés visés qui sont toujours les mêmes, entre horizontal et vertical. Exemples de marché sursaturés : la détection de fraudes dans la finance et et l'analyse prédictive du comportement des consommateurs dans le marketing en ligne et mobile.
- Aux USA, on y trouve parfois les ombres de la **DARPA**, de la **NSA** et de la **CIA** comme clients voire même comme investisseurs pour cette dernière, via son fonds **InQTel**. Surtout pour les solutions "horizontales". Ce n'est pas une question de "Small Business Act" mais simplement de besoins de ces organisations de défense et de renseignement !
- On retrouve aussi beaucoup d'anciens des Universités de **Stanford** et du **MIT** dans les startups de l'IA, généralement bardés d'un ou de plusieurs PhD en IA. Ils sont issus du monde entier.
- Une bonne moitié des startups deep tech crées aux USA l'ont été par des **immigrés**, provenant de Chine, Inde, Europe, Israël et ailleurs. Cette proportion avait baissé alors que l'Administration Trump rendait de plus en plus difficile l'installation d'étrangers, même qualifiés, aux USA. Reste à voir si l'administration Biden va relâcher cela.
- Les **technologies d'IA** employées sont assez mal documentées. Le machine learning et le deep learning reviennent souvent sans que l'on puisse évaluer si les startups ont réellement fait avancer l'état de l'art. Comme il se doit, une startup doit présenter un risque marché plus qu'un risque technologique ou scientifique. C'est pourquoi les startups de l'IA sont généralement positionnées dans l'application de techniques d'IA connues à des marchés divers, horizontaux ou verticaux. Elles profitent aussi parfois de l'effet d'opportunité en labellisant "IA" des projets qui quelques années auparavant auraient été vendus sous le sceau du "big data".
- Les solutions sont très souvent proposées sous la forme d'**APIs en cloud** mais les approches plateformes sont encore émergentes car elles ne bénéficient pas d'un effet push/pull courant dans le grand public (la demande pour des smartphones Android entrainant celles d'applications tournant dessus).
- Les **levées de fonds** sont de plus en plus élevées, avec de plus en plus de startups récoltant plus de $100M. Mais c'est peu par rapport aux décacornes qui ont des financements pouvant dépasser $1B. Les licornes sont d'ailleurs souvent des startups avec des offres grand-public.

[1704] Source : IDC FutureScape: Worldwide IT Industry 2017 Predictions.

Cartographie

Il existe de nombreuses sources de cartographies de startups de l'intelligence artificielle, notamment chez **CBInsights**, **VentureScanner** et **FranceIsAI**, pour le marché français dans ce dernier cas. Ces cartographies sont apparues vers 2015 lorsque l'investissement dans l'IA a commencé à décoller. La cartographie *ci-dessous* provient de CBInsights et date de février 2019[1705].

Grosso modo, on peut simplifier la segmentation des startups de l'IA comme suit avec :

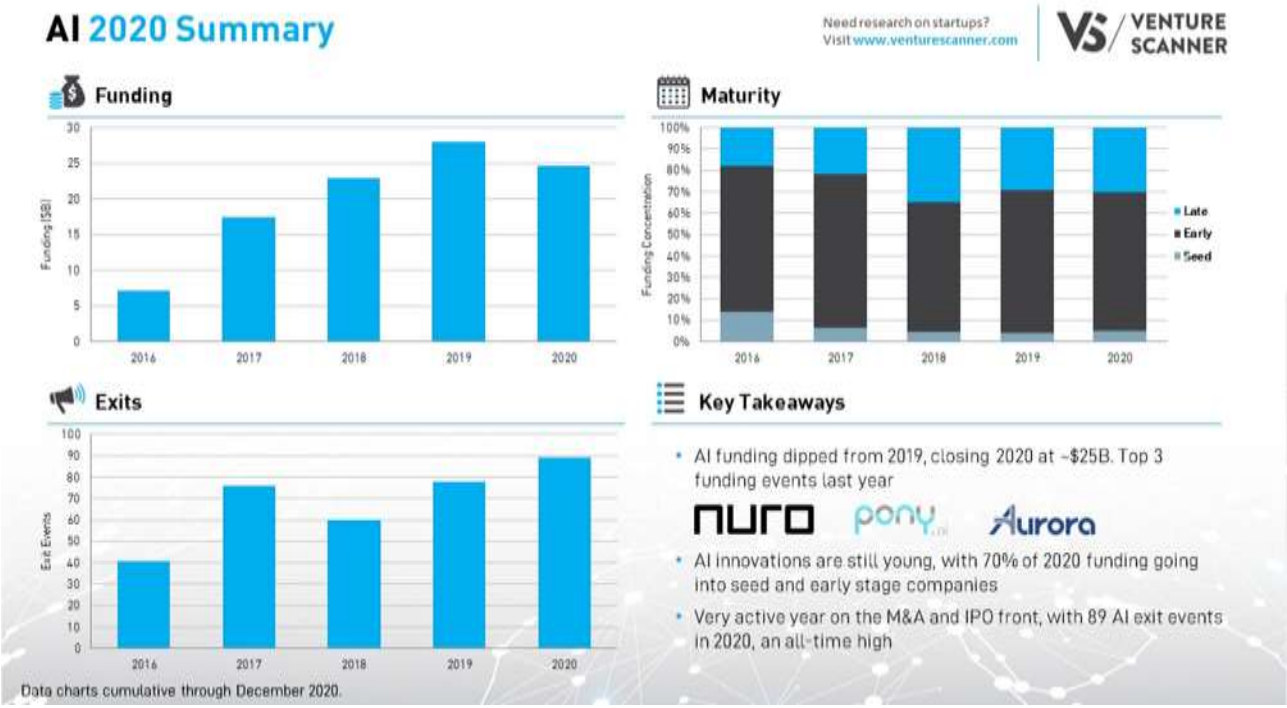
- **Plateformes** et **outils** : machine learning, deep learning, réseaux de neurones, composants. Ce sont les outils de base pour les développeurs et créateurs d'applications. Cela comprend aussi les outils de préparation et de manipulation de données à base de machine learning.
- **Applications horizontales** : avec les outils génériques liés au traitement de l'image et du langage, les outils de la cybersécurité, de l'IT (AIOps), pour les RH, la comptabilité et la finance des entreprises.
- **Applications verticales** : organisées par marché, transport, santé, finance, distribution, etc. Elles sont nombreuses mais c'est là que les éditeurs de logiciels et startups ont le plus de mal à trouver des économies d'échelle.

Dans la pratique, la frontière entre plateformes génériques et applications métiers est ténue. Nous avons par exemple vu dans le domaine de la santé des startups spécialisées dans l'imagerie médicale sur une seule pathologie (rétinopathie diabétique, dermatologie) ou d'autres qui en couvraient plusieurs.

[1705] Voir AI 100: The Artificial Intelligence Startups Redefining Industries, CBInsights, février 2019.

Dans le traitement des données, des sociétés de plateformes sont en fait rapidement spécialisées dans le marketing ou la finance. On constate une tendance baissière du financement des startups depuis 2017, dans le graphe *ci-dessous* issu de **Venture Scanner** fin 2020.

Comme si la vague était passée, laissant indiquer que « les jeux sont déjà faits » dans ce secteur d'activité. En Europe, environ 8% des startups financées sont dans l'IA.



Comme à chaque nouvelle vague technologique, celle de l'IA donne lieu à ce qui s'apparente à une bulle[1706]. Une bulle en nombre de startups créées et en montants levés, surtout aux USA et en Chine. Plusieurs startups ont réussi à lever plus de $200M, mais ces montants sont encore modestes au regard de nombreuses unicorns ayant un patrimoine technologique assez réduit au départ comme les Pinterest de ce monde.

Les acquisitions de startups suivent un bon rythme depuis 2017[1707]. Le financement des startups de l'IA s'apparente à une bulle pour au moins trois raisons.

Tout d'abord, nombre de ces startups n'ont pas une **scalabilité** évidente, surtout lorsqu'elles sont b2b. C'est moins vrai pour les plateformes qui sont les vendeurs de pelles et de pioches du marché mais chez qui la concurrence sera rude et fera émerger une toute petit poignée de leaders.

Cette difficulté à passer à l'échelle est amplifiée par l'IA car les données et les structures de données qui les alimentent sont souvent spécifiques à chaque client[1708].

[1706] Voir Pour un chercheur d'Oxford, l'intelligence artificielle est une bulle spéculative, avril 2018, qui relate les propos de Michael Wooldridge, le dirigeant du département des sciences informatiques de l'Université d'Oxford. Il considère que beaucoup de fantasmes sont plaqués sur les capacités de l'IA. Elle progresse plus lentement qu'il n'y paraît. Donc, les bénéfices de l'IA brandis par les startups sont survendus. Dans Why I don't invest in AI, juillet 2018, Peng Ong évoque le cas spécifique des startups d'IA s'attaquant à des marchés verticaux. La commoditisation de l'IA comme outil les rend rapidement moins différenciées. Les barrières à l'entrée sont similaires à l'habitude.

[1707] Voir The Race For AI: Here Are The Tech Giants Rushing To Snap Up Artificial Intelligence Startups, septembre 2019.

[1708] Voir The New Business of AI (and How It's Different From Traditional Software) par Martin Casado and Matt Bornstein, février 2020 qui reprennent ce point sur la scalabilité du modèle économique de nombreuses startups de l'IA.

Seules les startups b2b exploitant des données indépendantes des entreprises clientes peuvent véritablement scaler. Cela pourrait être le cas de startups telles qu'Equals 3 et d'autres cas vus sur le marketing. L'IA est encore un véritable artisanat du logiciel. On est encore très loin d'une véritable industrialisation des procédés.

La seconde est la **banalisation** de l'usage de l'IA par les startups. L'IA ne va bientôt plus être la spécificité de quelques startups avant-gardistes mais devenir courante chez presque toutes les startups et tous les éditeurs de logiciels.

N'importe quel éditeur de solution logicielle – cloud ou pas – qui s'alimente de données va pouvoir les exploiter pour faire du machine learning, de la segmentation et des prévisions.

La question se posera à chaque fois de la pertinence de le faire selon les volumes de données récoltés, mais on verra sans doute apparaître des méthodes permettant de se satisfaire de volumes de données limités. Bref, à terme, toutes les startups et tous les éditeurs de logiciels feront de l'IA et cela ne sera plus différenciateur à haut niveau !

S'y ajoute une autre tendance : nombre d'acquisitions de startups sont en fait des « *acquihires* [1709] », notamment chez Google. Des chercheurs ont tout intérêt à créer une startup avant de se faire recruter par Google.

Leur « sign-in » bonus sera ainsi plus élevé. Nombre d'acquisitions ont ainsi eu lieu pour des startups avec seulement un vague produit et pas vraiment de business model. C'est déjà arrivé par le passé mais le phénomène est plus marqué avec l'IA du fait de la rareté des compétences.

Évaluation

Quelles sont les caractéristiques d'une bonne startup faisant appel à de l'IA ? Ce sont quasiment les mêmes que pour les grandes entreprises évoquées dans la partie précédente. Ce petit inventaire est surtout destiné aux entreprises clientes mais est aussi utilisé par les investisseurs.

- **Talents** : une startup dans l'IA doit faire appel à des talents techniques variés avec des data-scientists, des développeurs maîtrisant le paramétrage de systèmes de machine learning et deep learning, et les autres connaissances techniques classiques allant du back-end au front-end. Le design est aussi de la partie car les solutions logicielles se distinguent encore par là. La première chose à faire pour évaluer une startup consiste donc à consulter la biographie de ses fondateurs et de ses équipes techniques[1710].
- **Métier** : une startup connaît bien le métier de ses clients, ses contraintes, ses besoins et aspirations. Elle sait créer une solution qui s'intègre bien dans l'existant. Elle sait parler métier aux décideurs de l'entreprise.
- **Données** : il n'y a pas d'IA sans accès à des données pour entraîner ses modèles. Les données exploitées par la startup peuvent être de trois types : ouvertes et facilement disponibles sur In-

[1709] Acquisitions based on hiring. Ce sont des acquisitions de startups qui sont faites pour recruter les développeurs et autres équipes techniques de la startup plus que pour récupérer un produit, leurs clients ou un marché.

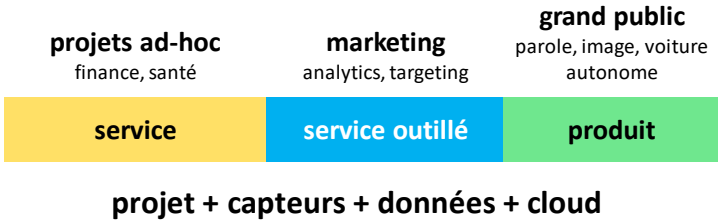
[1710] Voir How To Create An AI (Artificial Intelligence) Startup par Tom Taulli, 2018.

ternet[1711] (open data, ImageNet, WordNet, MNIST, TIMIT pour la reconnaissance de la parole en américain), collectées de manière exclusive à la startup via ses propres solutions matérielles et/ou logicielles, par exemple via ses objets connectés, ou provenant des systèmes d'information de ses propres clients. La différenciation de la solution provient généralement de la combinaison des trois sources. Une startup n'exploitant que des données ouvertes aura moins de barrières à l'entrée.

Et pour accéder aux données des entreprises clientes, il faudra souvent faire du spécifique ce qui réduira les effets d'économie d'échelle de la startup. Autre question clé : où sont stockées les données ? Comment est géré le respect de la vie privée des utilisateurs pour les applications grand public ? La startup est-elle conforme à la RGPD en vigueur depuis le 25 mai 2018 ?

- **Produit** : est-ce que la solution est générique ou demande-t-elle d'adopter un mode projet lourd pour sa mise en œuvre chez chaque client ? Si on est en mode projet à chaque fois, on sera dans la catégorie des services outillés, la startup étant hybride entre startup produit et entreprise de services du numérique (ESN) avec peu d'économies d'échelle[1712]. Et aussi, ne pas oublier d'avoir une démonstration du logiciel !

  Dans l'IA, l'ergonomie est aussi importante que la fonction ! Et s'il n'y a pas de produit, c'est que l'on n'a pas affaire à une startup[1713] ! La création d'un produit requiert une certaine capacité à se focaliser[1714].



- **Technologies** : quand une startup indique avoir créé « son IA », il est bon de creuser un peu pour se faire expliquer le pourquoi du comment. Quels outils a-t-elle exploité pour créer sa solution ? Quelles méthodes de machine learning ou quels types de réseaux de neurones ? Est-ce que leur combinaison est originale, ce qui créera une différenciation, que l'on devra retrouver dans la performance de la solution ? Quelle est la partie algorithmique qui est spécifique à la startup ? Est-ce qu'elle a développé un savoir-faire spécifique dans l'assemblage de briques algorithmiques diverses ?
- **Business** : quel est le modèle économique de la startup ? Est-il récurrent ? Où sont les économies d'échelle ? Des questions assez classiques.

[1711] Voir la liste de jeux de données d'entraînement sur Wikipedia sachant qu'elles ne sont pas toutes publiées en open data : https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research#Speech.

[1712] Element.ai (2016, Canada, $102M) est une startup de Montréal qui se focalisait sur la création de produits pour différents marchés verticaux comme la finance et la supply chain, en faisant le pont entre la recherche et les entreprises clientes. Leur équipe de fondateurs comprend le fameux chercheur Yoshua Bengio du MILA. Leurs investisseurs comprennaient Microsoft Ventures, Intel Capital et Nvidia. J'ai pu croiser son fondateur Jean-François Gamier à la résidence de l'ambassadeur du Canada à Paris le 10 octobre 2018. L'idée de la société était de créer des solutions génériques multi-clients architecturées autour de l'IA exploitant aussi bien des données sructurées et non structurées. Il reconnaissait que la création de produits était difficile en mode b2b et qu'il n'existait pas véritablement de startup « best in class » dans le domaine. L'un des manières de procéder est de trouver des idées communes à plusieurs clients que les clients sont prêts à adopter de manière non exclusive. La startup a déjà embauché 40 développeurs Français. Ils ont même des grands clients en France. Mais elle a été revendue à Service Now en 2020 comme déjà vu ailleurs dans ce document.

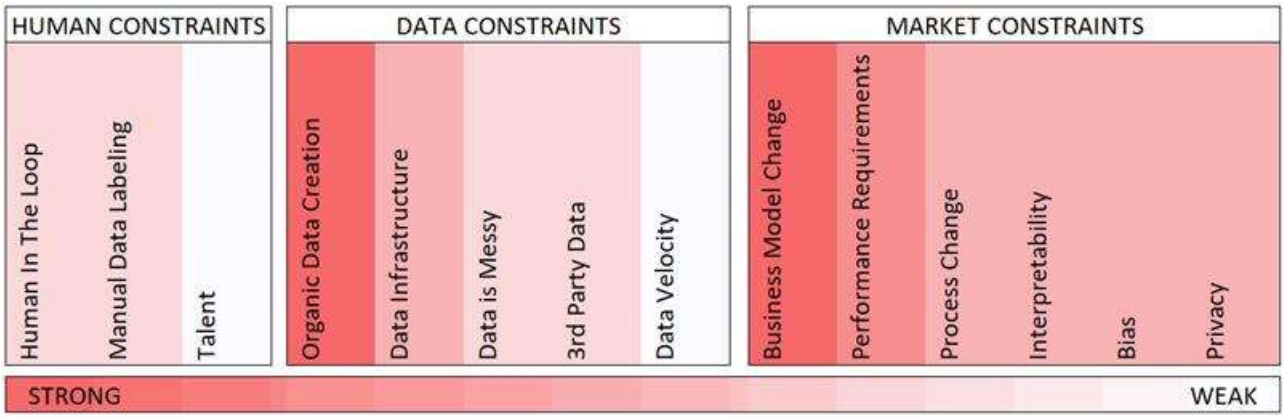
[1713] De ce point de vue-là, la levée de fonds de 2M€ de Nabla.ai, créée par Alexandre Lebrun de Facebook est étonnante dans la mesure où la société se positionne comme une société de services. La levée s'explique par le pédigrée du fondateur qui avait déjà vendu une première startup à Facebook, wit.ai. Voir Nabla, la pépite IA créée par des anciens de Facebook et de MyLittleParis, 2018.

[1714] Je suis tombé un jour sur une startup dont je tairais l'identité mais qui se reconnaîtra peut-être, et qui voulait créer à la fois une plateforme de création de chatbots et un moteur de machine learning. Il s'agissait visiblement plutôt d'une société faisant du service outillé qu'une véritable startup produit pouvant faire avancer l'état de l'art dans un de ces deux domaines.

- **Financement** : c'est toujours le nerf de la guerre pour le développement de véritables startups, celles qui ont une grande ambition, notamment internationale. Nous avons vu dans les énumérations nombreuses de ce document que les startups US bénéficiaient souvent de financements importants, pouvant facilement dépasser les $30M, ce qui est plus rare en France et en Europe en général.

La liste ci-dessus est mise en forme différemment dans une production d'un certain AJ Christensen qui décrit les défis que les startups de l'IA doivent surmonter et sous-estiment généralement[1715].

Au top des problèmes à résoudre : la qualité des données et le modèle économique[1716].



Ecosystème français de l'IA

Dans tous les nouveaux domaines technologiques, la France aspire collectivement à l'excellence et à la singularité. Pas celle de l'IA mais celle de la différenciation.

L'habitude est de mettre en valeur voire de monter en épingle l'excellence de nos chercheurs puis de nos ingénieurs, qu'ils soient restés en France ou expatriés dans de grandes entreprises internationales du numérique. Cela doit justifier notre place au soleil de la compétitivité scientifique, technologique et économique. Et nous avons Yann Le Cun, le père des réseaux de neurones convolutifs, qui œuvre chez Facebook, et Luc Julia, père de Siri, chez Samsung, etc !

Depuis 2016, diverses initiatives cherchent à valoriser les startups de ce nouveau marché et de mettre en avance l'excellence française et ses opportunités. Nous avons eu notamment le plan France IA du gouvernement en 2017, le rapport de la Mission Villani en 2018, la nomination d'un coordinateur de l'IA au sein de la DINSIC, Bertrand Pailhes en juillet 2018 (qui quittait ce poste mi novembre 2019) et la création de l'association France Is AI d'ISAI en 2016, rattachée depuis à France Digitale.

Ce comportement est assez fréquent face à de nombreuses vagues technologiques : dans les jeux vidéos, dans les objets connectés, dans la cybersécurité, la robotique, et récemment, pour la Blockchain pour ne prendre que quelques exemples. Le timing de ces initiatives est variable. Il est en tout cas relativement tardif pour ce qui est de l'IA.

L'IA constitue bel et bien une opportunité de positionner le pays, à la fois ses entreprises numériques et les autres qui peuvent faire appel à l'IA pour innover et devenir plus compétitives. Mais nous sommes aveuglés par les mêmes erreurs de perspective que par le passé.

[1715] Voir aussi les recommandations de 10 Rules Entrepreneurs Need to Know Before Adopting AI par Rocio Wu, février 2020. Avec entre autres : comprendre le besoin client est plus important que les méthodes utilisées pour les traiter, développer une stratégie d'acquisition de données dès le départ, penser à l'interface utilisateur et enfin, recruter des ingénieurs en IA mais aussi des experts du domaine métier.

[1716] Voir Don't Bet on AI (yet) I've analyzed 7,000 "AI Startups". Most underestimate the challenges that plague AI. Does yours? par AJ Christensen, juillet 2019.

Ce qui fera les forces et les faiblesses de ces startups n'est pas directement lié à l'IA mais plutôt générique. L'important est de savoir à quelle vitesse ces startups se financent, développent un prototype viable, génèrent leurs premières références clients et se déploient commercialement à l'étranger, surtout aux USA. Ces besoins sont génériques.

L'IA a cependant quelques caractéristiques qui ralentissent ces élans : les économies d'échelle sont bien plus grandes dans les activités tournées vers le grand public que vers les entreprises. Vers ces dernières, les startups de l'IA ont du mal à générer des économies d'échelle car chaque entreprise génère son propre projet avec ses spécificités et ses données propriétaires.

Les économies d'échelle ne fonctionnent bien qu'avec des produits répondant à des besoins génériques, avec des plateformes de développement ou dans une certaine mesure avec des outils d'IA qui la démocratisent à une plus large audience, par exemple les TPE et PME.

Dans l'IA comme ailleurs dans le numérique, l'intégration est mondiale, pas française. Une startup ne peut devenir un leader mondial qu'en devenant une plateforme, c'est-à-dire, un produit extensible par des tiers. Les plateformes ne sont pas françaises pour le marché français, elles sont mondiales.

L'avantage des startups américaines est dans la dimension de leur marché intérieur, qui conditionne à la fois leur surface commerciale initiale et leur capacité de financement. Pour les égaler, il faut trouver cette surface. Or la surface française est toujours comprise entre 1/7,8 (PIB) et 1/32 (financement dans le capital risque) vis-à-vis des USA. Il faut donc voir grand pour les meilleurs.

Cela correspond d'ailleurs à un progrès récent de l'écosystème entrepreneurial français. On compte chaque année une trentaine de startups qui dépassent les 10M€ de financement. Et elles s'orientent de plus en plus à l'international. Il faut continuer. L'investissement dans les startups atteignait 4Md€ en 2018, un niveau extraordinaire par rapport au 1Md€ qui sévissait il y a à peine quelques années. Ils atteignaient même 5,4Md€ en 2020, en pleine période de pandémie. Mais ces investissements ont aussi grandi dans nombre d'autres pays. Dans l'ensemble de ce document, j'indique systématiquement les montants levés par les startups citées. Cela calme parfois un peu !

Le timing d'une mobilisation autour d'une nouvelle technologie définit indirectement le résultat que l'on peut en obtenir. Comme les principales plateformes du marché sont américaines, on devient des usagers de ces plateformes. Les économies d'échelle vont aux américains et le service et la personnalisation à nos acteurs.

Qui plus est, l'IA va très rapidement devenir une commodité. Dans quelques années, il n'y aura plus de marché de l'IA. Il se confondra avec celui de l'informatique et du numérique, tant dans le grand public que dans les entreprises. La question clé sera alors son adoption par l'ensemble des entreprises et bien moins dans la création d'une industrie numérique de l'IA où les jeux auront été faits. On parlera alors de rattrapage des TPE et des PME, comme on a dû le faire – et pas terminé – pour Internet.

Le développement de l'écosystème de l'IA, moins visible, concernera alors les prestataires de services et ceux qui proposent du service outillé. Nombre d'agences de communication, web agencies et autres entreprises de services numériques se mettent progressivement à l'IA et structurent leurs offres.

Recherche

Faisons d'abord un tour dans l'écosystème de la recherche en IA français. Il a été très bien inventorié dans le cadre du plan **France IA** du gouvernement, publié en mars 2017[1717].

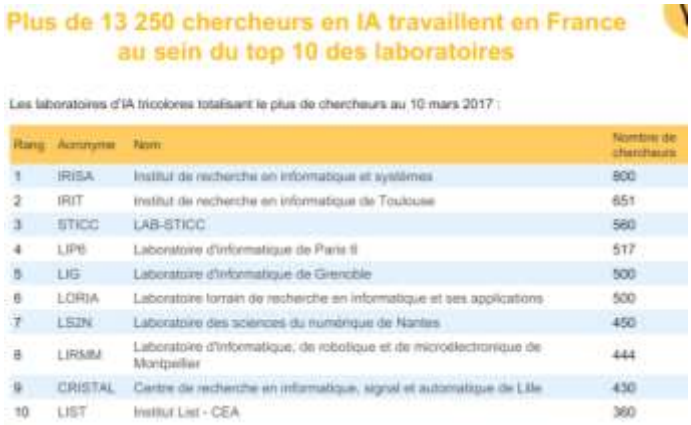
[1717] Voir https://www.economie.gouv.fr/France-IA-intelligence-artificielle. Ce plan a été lancé par Axelle Lemaire puis repris par son successeur Christophe Sirrugue. Il faut considérer que le plan présenté en mars 2017 était un rapport d'étape. Après la présidentielle, le nouveau gouvernement a confié à Cédric Villani la mission de créer un plan complet, qui a été présenté fin mars 2018

Il a probablement dû évoluer depuis, mais à la marge. Selon ce rapport, la recherche publique française serait disséminée dans plus de 220 équipes de recherche totalisant 5300 chercheurs, avec de nombreux projets collaboratifs associant laboratoires publics, universités et, parfois, entreprises privées. Les principaux organismes se focalisant sur l'IA sont Inria, le CNRS et le CEA, notamment le LIST. On peut y ajouter l'IRT **b<>com** situé en Bretagne qui explorer de nombreux domaines de l'IA, dans le traitement de l'image et du langage, dans la santé et dans les usages liés aux télécoms. La recherche en IA est soit « pure », soit appliquée à des domaines comme la santé.

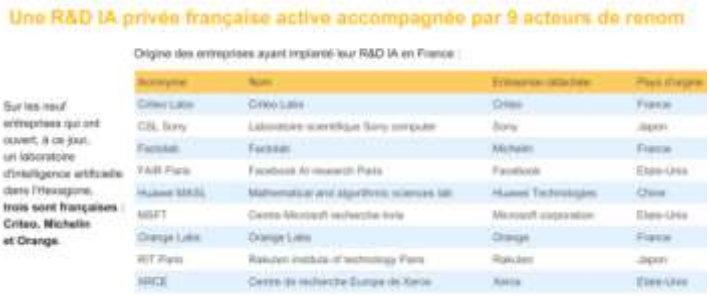
Les chercheurs français seraient les plus prolixes en publications scientifiques, derrière les Américains et les Chinois qui dominent le secteur. Les Canadiens ne sont pas loin, aussi bien à Toronto qu'au Québec. En fait, la France serait 7e en publications scientifiques dans l'IA en 2018[1718], derrière UK, l'Allemagne, l'Inde, le Japon et bien évidemment, la Chine et les USA, talonnée de très près par l'Italie et l'Espagne, mais devant le Canada. Pourtant, dans nombre d'inventaires de la presse anglo-saxonne, la France est royalement ignorée[1719].

Nos politiques aiment à rappeler l'excellence de la recherche française[1720]. Mais quels sont les indicateurs pertinents pour évaluer cette excellence ? Il existe plusieurs manières de classer les pays au niveau des publications scientifiques.

Selon une source de la même époque, mars 2017[1721], la France compterait en fait 13 250 chercheurs en IA. C'est probablement une estimation un peu élevée. L'évaluation du nombre de ces chercheurs n'est pas évidente. Est-ce qu'un chercheur qui utilise de l'IA dans ses travaux dans la santé est un chercheur « en » IA ? Ce serait comme si on considérait qu'un chercheur en sciences des matériaux utilisant un ordinateur était un chercheur en logiciels.



La recherche privée en IA en France serait dominée par les laboratoires ouverts par des entreprises étrangères : Sony, Facebook, Huawei, Microsoft, Rakuten et Xerox. Sachant qu'ils ne sont pas dédiés à 100% à de la recherche en IA. Il faudrait y ajouter le laboratoire d'IA installé à Grenoble par le Coréen Naver pendant l'été 2017.



Ce laboratoire occupe 70 chercheurs à comparer aux 200 chercheurs de Naver situés en Corée du Sud. Naver avait repris clé en main l'ancien centre de recherche de Xerox de Grenoble créé en 1993.

[1718] Voir https://www.scimagojr.com/countryrank.php?year=2017&category=1702. La France est également au septième rang sur la période 1996-2017. Cependant, en 2012 et 2013, l'Espagne était devant la France.

[1719] Voir une petite revue de presse des pays leaders de l'IA qui ignore royalement la France : The AI race: 7 countries to watch out for, février 2018, 2017 in Review: 10 Leading AI Hubs, décembre 2017, 5 Countries Leading the Way in AI, janvier 2018, These Seven Countries Are In A Race To Rule The World With AI, décembre 2017 (qui cite l'Estonie et la Russie, mais pas la France !).

[1720] Voir Intelligence artificielle : "Les atouts de la France sont multiples" dont "l'excellence de la recherche française", avance Cédric O par France Info, mars 2020.

[1721] Voir Intelligence artificielle en France : la carte des laboratoires de Lélia De Matharel du JDN, mars 2017. Ces 13250 chercheurs ne sont pas mentionnés dans la page recherche du site France Is AI. Selon le Ministère de l'Enseignement Supérieur et de la Recherche, il y aurait environ 428 600 équivalents temps plein chercheurs en France en 2017.

Le laboratoire est spécialisé en vision artificielle, mobilité et traitement du langage. Naver a aussi lancé un incubateur à Station F où il héberge notamment Videolabs, et a financé à hauteur de 200M€ le fonds d'investissement Korelya Capital lancé par Fleur Pellerin.

Dans le privé, Criteo[1722], Orange[1723] et Michelin sortent du lot. Inria avait publié courant 2016 un excellent livre blanc qui décrivait ses priorités et projets dans l'IA[1724]. De manière assez classique, les projets portaient sur le langage, la vision et la robotique. Inria planche aussi beaucoup sur l'IA symbolique avec le web sémantique, les neurosciences et sciences cognitives ainsi que sur la programmation par contrainte. Elle s'intéresse à la protection de la vie privée ainsi qu'aux applications de l'IA dans la santé.

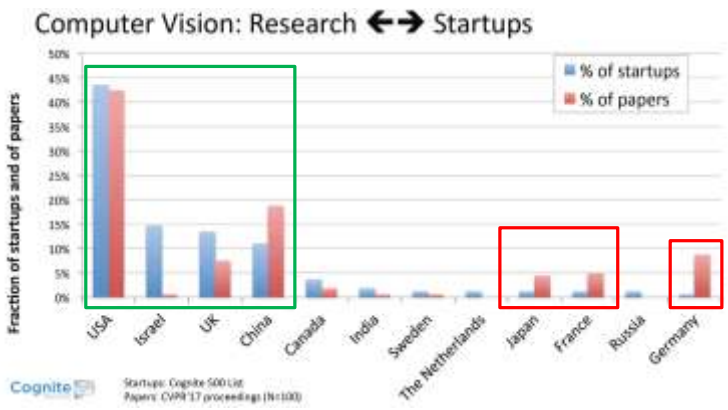
Il est difficile de caractériser les spécificités de la recherche française en IA. Elle est multi-domaine sans spécialisation apparente.

On peut cependant y distinguer une force dans l'IA symbolique et la logique formelle, dans le traitement du langage ainsi, en filigrane, qu'un souci de créer des solutions d'IA responsables et éthiques[1725].

Le défi pour ces chercheurs et leurs autorités de tutelle est de trouver des applications marchés de leurs travaux. En consultant la liste des participations d'**IT-Translation**[1726] qui est l'un principaux financeurs de projets issus d'Inria, on constate que l'IA est souvent en filigrane de ces projets. Les travaux des chercheurs en IA n'aboutissent pas naturellement à des projets entrepreneuriaux. Ne serait-ce parce qu'il faut une couche de traduction entre ces réalisations et leurs applications et que les innovations des startups résultent souvent de la combinaison de plusieurs méthodes et techniques[1727].

Cognite Ventures faisait un point en 2017 de la création de startups dans un domaine précis, la vision artificielle, au regard des publications scientifiques dans le domaine[1728].

Il fait ressortir deux types de pays : ceux qui valorisent bien leurs travaux de recherche dans des startups (USA, UK, Chine) et ceux qui au contraire publient bien plus qu'ils ne créent de startups (France, Japon, Allemagne).



[1722] En juin 2018, Criteo annonçait investir 20M€ sur trois ans pour créer son laboratoire de recherche en IA à Paris, qui complète ses équipes produits qui comprennent déjà 300 ingénieurs et développeurs. Comme il se doit, ce Criteo AI Lab sera très focalisé sur le machine learning. Le laboratoire est dirigé par une chercheuse indienne, Suju Rajan, que l'on voit fréquemment intervenir dans les conférences en France.

[1723] Orange a toujours eu un gros laboratoire de recherche. Actuellement dirigé par Nicolas Demassieux et situé principalement dans les locaux Orange Gardens à Chatillon sous Bagneux, avec quelques autres équipes à Rennes et Lannion, il comprend diverses équipes qui exploitent des briques d'IA dans leurs recherches. Cela porte notamment sur des aspects très opérationnels du déploiement à venir des réseaux 5G.

[1724] Voir Intelligence artificielle, dédis actuels et l'action d'Inria, 2016 (82 pages).

[1725] Cela se retrouve notamment dans les travaux de Laurence Devillers, du CNRS-LIMSI qui portent sur le langage, sur la détection des émotions et sur l'éthique de l'IA.

[1726] Voir le portefeuille de participations d'IT Translation.

[1727] Pour apprécier la difficulté à transformer la recherche en IA en solutions métiers, vous pouvez par exemple consulter les actes de la dernière conférence ICML sur le machine learning qui s'est tenue début août 2017 en Australie : https://2017.icml.cc/Conferences/2017/Schedule.

[1728] Voir Does vision research drive deep learning startups?, août 2017. Le chart de cette page ne correspond pas exactement à celui de l'article. Je n'arrive pas à retrouver sa source.

Israël est un cas avec très peu de publications et beaucoup de startups. On peut difficilement extrapoler cela mais force est de constater que les startups issues de la recherche ne sont pas très nombreuses en France dans l'IA.

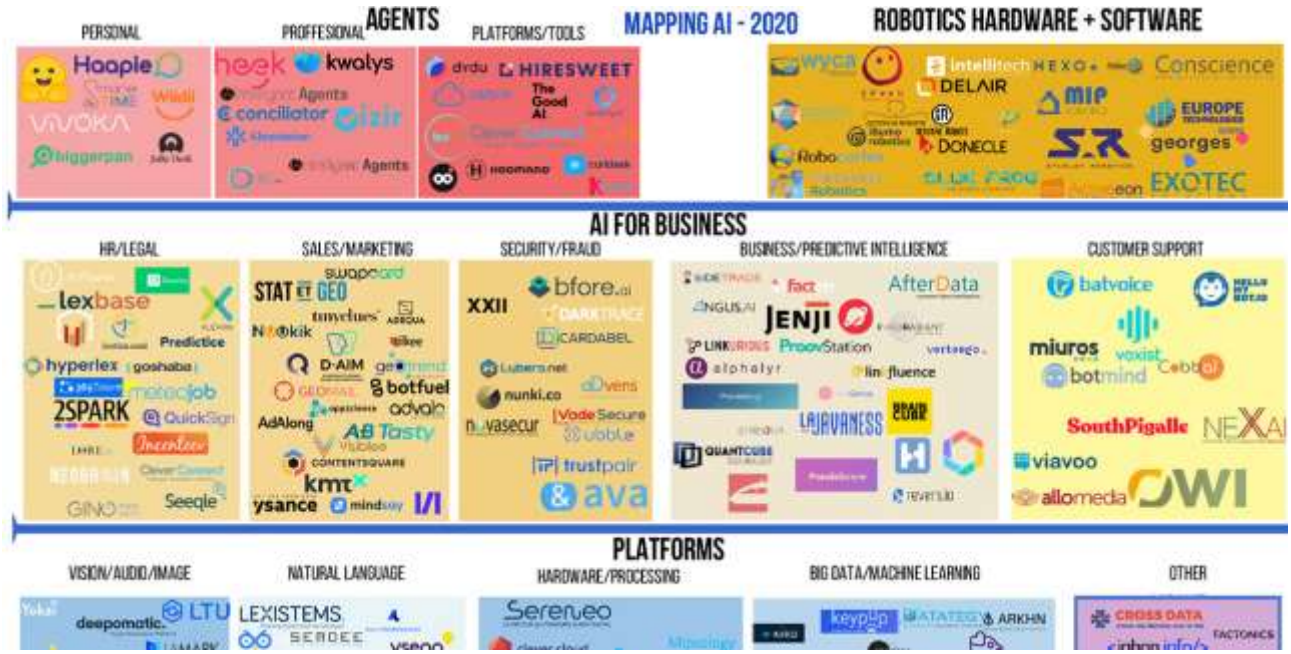
En France, la recherche dans l'IA semble mieux financée côté civil, même s'il est difficile de le vérifier par les chiffres. On ne s'en plaindra pas. A ceci près que la R&D militaire US a une qualité : elle est orientée vers des objectifs pratiques selon des cahiers des charges. De son côté, la recherche civile française fonctionne plutôt de manière très décentralisée.

Startups

Selon une étude de Roland Berger publiée en octobre 2019 pour le compte de l'association **France Digitale**, il y avait à cette date 432 startups de l'IA en France (pour 305 un an avant)[1729].

La France se retrouve en seconde position européenne selon les critères habituels de nombre de startups et de levées de fonds et ex-aequo avec le Royaume-Uni pour ce dernier point, pour $1,4B de levés entre 2014 et 2019[1730].

Le montant bénéficie de très grosses levées comme les $200M de **Meero**. Selon Bpifrance, le nombre de startups françaises de l'IA était de 552 dans une évaluation datant d'avril 2019, ayant levé 2Md€. Et 453 pour FranceIsAI/France Digitale en novembre 2020[1731].



La diversité de leur activité est très grande avec deux marchés verticaux qui émergent, celui de la santé et celui du retail. Elles sont à dominantes b2b. On y trouve un grand nombre de chatbots avec une difficulté à distinguer les sociétés de services des véritables plateformes technologiques de création de chatbots[1732].

[1729] Source du chart : https://franceisai.com/startups, octobre 2019. La liste est déclarative. Les startups s'inscrivent d'elles mêmes. Avec deux écueils : des startups qui n'utilisent pas forcément de briques logicielles d'IA ou celles qui sont des prestataires de services en mode projet et n'ont pas de produit, et tout au plus une boîte à outils plus ou moins propriétaire. Voir France : les startups spécialisées en IA ne cessent d'augmenter par Geneviève Fournie, octobre 2019.

[1730] Voir La France en deuxième position des pays européens les plus actifs dans l'IA par Antoine Crochet-Damais, novembre 2019.

[1731] Voir Les startups spécialisées dans l'IA continuent à augmenter en France, novembre 2020.

[1732] Voici un petit inventaire de startups de chatbots françaises : AskHub, Askr.ai, Botfuel, Botmind, Clustaar, Golem, Haapie, Heek, Hellomybot, Hubware, Opla, Owi, Posos, Smartly, Snapdata, Synapse Développement, VisualBot.ai, Voxist, Wiidii et xBrain. Ca fait du monde !

Entreprises

Nous en avons cité quelques-unes dans la grande partie de cet ebook qui est dédiée aux marchés verticaux : nombre de grandes entreprises ont adopté l'IA dans leur informatique interne. C'est plus visible dans les entreprises qui gèrent de gros volumes de données comme dans les télécoms, la finance, l'assurance, la distribution et la santé. Les utilities ont aussi fort à faire avec l'IA pour optimiser et maintenir leurs infrastructures.

Dans les transports, la plupart des acteurs, des constructeurs aux opérateurs, peuvent aussi faire appel à l'IA. Il en va de même de l'ensemble du complexe militaro-industriel.

Il est cependant peu évident de jauger de la vitesse à laquelle ces entreprises adoptent l'IA. Je n'ai pas encore trouvé de benchmark international. En gros, les équipes innovation et les entreprises les plus hardies sont déjà à plein pied dans l'IA. Vu de près, la majorité des grandes entreprises ont au moins déployé une douzaine de projets à base d'IA. En 2019, on est passé des PoC au retour d'expérience en phase de production et parfois avec des désillusions comme pour les chatbots qui sont souvent décevants.

Associations

La France est le pays des associations loi 1901 et cela commence à se refléter dans les associations autour de l'IA.

A commencer par l'**Association Française pour l'Intelligence Artificielle** (AFIA), créée en 1993, et qui fait la promotion de la recherche en IA. Elle regroupe donc surtout des chercheurs. Comme il y a toujours du pour et du contre, nous avons aussi une **Association contre l'IA**, l'AFCIA, créée en 2015 et qui vise simplement à interdire à l'échelle mondiale toute recherche sur l'IA[1733]. C'est l'équivalent anti-OGM ou anti-nucléaire appliqué à l'IA.

On retrouve une dichotomie avec deux associations : d'un côté l'initiative **FranceIsAI** de France Digitale et de l'autre, le **Hub France IA**, deux mouvements qui ont des visées différentes. **FranceIsAI** est intégrée dans l'association France Digitale qui associe entrepreneurs et investisseurs dans les startups françaises. Elle produit un inventaire des startups en France régulièrement mis à jour (cf. page précédente) et organise notamment l'excellente conférence FranceIsAI chaque année à l'automne, qui met en valeur à la fois les chercheurs et les entrepreneurs du secteur.

Le **Hub France IA** a été créé par des chercheurs et industriels de l'IA français. Il a pour vocation de fournir à l'écosystème une plateforme de ressources visant a stimuler l'accélération de l'adoption de l'IA par les différents acteurs de l'économie, surtout les entreprises, notamment via des groupes de travail thématiques, associant grands groupes, startups et laboratoires de recherche en IA, une offre d'initiation et de formation à l'IA, des rencontres et du networking entre membres et des aides à la constitution de dossiers de financement publics pour des programmes de R&D.

Dans le même genre, l'**IMA** (Innovation Makers Alliance) regroupe plus de 60 entreprises françaises sur l'IA, promeut l'innovation deeptech. Son groupe de travail sur l'IA a été fondé en 2017.

Il existe aussi une association l'**IA pour l'École**, qui vise à créer des ponts entre les pionniers de l'IA et tous les acteurs de l'éducation.

Un point clé : l'IA ne doit pas être qu'une affaire d'hommes. Comme dans le développement logiciel, on y trouve malheureusement plutôt une minorité de femmes alors que ces technologies vont conditionner le futur de l'humanité et du travail. Pourtant, on trouve plein de femmes remarquables dans l'IA, comme le montre cet inventaire US[1734]. D'où l'intérêt d'initiatives telles que **Women in AI**, une association mondiale avec une branche en France qui fédère les femmes travaillant dans le

[1733] Voir On peut être contre l'intelligence artificielle par principe de Irénée Régnauld, publié sur Uzbek&Rica en janvier 2017

[1734] Voir Meet these incredible women advancing AI research, Topbots, mai 2017.

secteur de l'intelligence artificielle et qui cherchent à attirer d'autres femmes dans le domaine. Des initiatives communautaires ont été également lancées en région comme **Lyon in AI**, piloté par Amélie Cordier, ancienne CTO de la startup de logiciels en robotique **Hoomano**.

Partenariats internationaux

En juin 2018, Emmanuel Macron et Justin Trudeau lançaient un projet franco-canadien en IA, en l'espèce, un groupe international d'étude réunissant des experts indépendants (scientifiques, gouvernements, industrie et société civile). Il s'agit visiblement surtout d'un groupe de réflexion sur la régulation de l'IA, pas vraiment d'un laboratoire de recherche. Le principal livrable ? La création d'un groupe de travail qui devra définir les contours de ce projet. Cela se concrétisait un tant soit plus fin octobre 2019, avec l'annonce par Emmanuel Macron lors de son discours à la fin du forum **AI for Humanity** de la création de deux centres d'expertise sur l'éthique de l'IA en 2020 à Paris et Montréal.

Enfin, une quinzaine de pays dont la France et le Canada lançaient en juin 2020 le **GPAI** (Global Partnership on AI, Partenariat Mondial pour l'Intelligence Artificielle en français[1735]), autre initiative destinée à favoriser le développement d'une IA fondée sur les droits de l'Homme, une IA responsable, l'inclusion, la diversité, tout en favorisant l'innovation et la croissance. Les pays impliqués s'engagent à soutenir la recherche et les usages allant dans ce sens. Le GPAI s'appuie sur un groupe de travail constitué d'experts autour de quatre thèmes : l'utilisation responsable de l'IA, la gouvernance des données, l'avenir du travail et l'innovation. Le PMIA est géré par l'OCDE dont le siège est à Paris ainsi que dans deux centres, à Paris à l'Inria et à Montréal. Le projet associe aussi l'UNESCO.

L'OCDE a aussi son **AI Policy Observatory** (OECD.AI), lancé en février 2020, une plateforme en ligne pour les états, les chercheurs et autres et destinée à partager des informations servant à définir des politiques publiques de l'IA. C'est un peu l'équivalent de l'EuroStat de l'IA. Cela complète les principes de l'IA de l'OCDE (OECD AI Principles) qui ont été adoptés par 42 pays en mai 2019. L'OCDE est aussi impliquée dans GPAI.

On peut aussi citer l'**AI Rome Call for AI ethics** lancé en 2020 avec l'appui du Saint Siège ainsi que de... Microsoft et IBM. Le mélange de la carpe diem et des lapins ! Tout cela pour ressasser les mêmes objectifs que dans la déclaration de Montréal : faire des IA transparentes et explicables, créer des solutions inclusives, promouvoir la responsabilité, l'impartialité, la fiabilité et préserver la sécurité et la vie privée.

Il existe une autre alliance, **GAIAA** (Global Artificial Intelligence Academic Alliance), à ne pas confondre avec GAIAX, l'initiative de cloud européenne lancée en 2020. Elle associe des universités américaines, chinoises et australiennes. Lancée en 2018, elle vise à développer la recherche collaborative mondiale en IA. Elle comprend notamment la startup chinoise **SenseTime** qui est spécialisée en vision artificielle, notamment pour la vidéo surveillance.

Les acronymes en GxxxA ne vont pas s'arrêter là puisqu'une certaine Amy Webb propose depuis 2019 de lancer une **GAIA** (Global Alliance on Intelligence Augmentation) pour créer des normes sur l'IA qui permettraient de résister à l'emprise de la G-MAFIA, une variante des GAFA qui ajoute Microsoft et IBM (au lieu de GAFAMI)[1736].

Que dire ? Trop de comités vont tuer les comités !

[1735] Avec le Canada, l'Allemagne, l'Australie, la Corée du Sud, les USA, l'Italie, l'Inde, le Japon, le Mexique, la Nouvelle-Zélande, le Royaume-Uni, Singapour, la Slovénie et l'Union européenne.

[1736] Voir Why AI is a threat to democracy—and what we can do to stop it par Karen Hao, février 2019 qui porte sur le livre The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity d'Amy Webb publié en mars 2019 (302 pages). Cette dernière dénonce de manière un peu manichéenne la concentration du pouvoir qui est entre les mains des GAFAMI. Son avertissement perd une bonne part de crédibilité lorsqu'elle le construit autour d'un inéluctable l'avènement de l'AGI.

IA et société

Nous allons sortir ici des considérations techniques et d'entreprises pour aborder la place de l'IA dans la société et tenter de traiter quelques questions clés.

Quels bouleversements annonce-t-elle, notamment dans le travail et l'emploi ? Quelles craintes et espoirs soulève-t-elle ? Comment la politique et l'Etat s'en emparent-ils ? Comment la réglementation pourrait-elle évoluer ? Quel est le rôle des entreprises de ces points de vue-là ? Pourquoi parle-t-on d'éthique de l'IA ? Que penser des prophètes de la singularité ? C'est l'objet de cette partie.

Craintes sur l'IA

L'IA génère-t-elle plus de craintes que les machines à tisser, les chemins de fer, l'aviation commerciale, l'énergie nucléaire ou les OGM au moment de leur apparition ? Il est difficile de comparer des époques différentes mais l'IA est en tout cas entrée dans le club plutôt fermé des technologies numériques qui font plus peur que la moyenne.

Une bonne part de ces peurs provient de la science fiction plus que de la science, ainsi que des effets d'annonce enjolivés à la moindre avancée du deep learning et à une conception étriquée de la notion même d'intelligence humaine. On use et on abuse trop facilement de la loi de Moore, simplifiant à volonté la notion même d'intelligence humaine pour la comparer à celle des machines. On confond métiers et tâches, calculs et décisions, automatisation et outillage. Et le moutonisme médiatique fait son effet, comme la titraille ci-dessous l'illustre.

Et cela se retrouve dans les enquêtes d'opinion. Selon l'enquête iLife de l'agence de communication BETC du groupe Publicis réalisée en février 2017 auprès de 12 169 personnes de plus de 18 ans, les français sont plutôt pessimistes sur l'impact de l'IA sur la société, sur l'emploi et sur la vie en général[1737].

L'IA conserve un côté magique qui permet de faire prendre des vessies pour des lanternes à de vastes audiences, y compris celles qui sont éduquées. Mais, même en étant prudent et conservateur, on peut estimer que l'IA aura un impact aussi important que les 35 années de vagues d'innovations numériques qui viennent de se succéder.

[1737] Voir l'étude iLife de BETC (slides). L'IA est bonne pour la société : Chine : 89%, USA : 53%, France : 33%. L'IA va créer des emplois : Chine : 59%, USA : 48%, France : 34%. L'IA impactera positivement nos vies : Chine : 45%, USA : 54%, France : 64%. L'IA nous libérera : Chine : 78%, USA : 46%, France : 36%.

C'est au minimum une grande vague de « logiciels 2.0 » qui est lancée à vive allure. Mais au même titre que la situation d'aujourd'hui était difficile à prédire il y a 35 ans, celle des 35 ans l'est tout autant. La raison est simple : les progrès de l'IA sont non seulement incertains d'un point de vue scientifique et technologique, mais s'y ajoutent les habituelles dimensions économiques et sociales qui jouent toujours le rôle de filtre entre l'univers du possible et ce qui devient disponible.

Une bonne part des craintes provient aussi de la propension à projeter sur les robots et l'IA nos propres travers. La vision antropomorphique de l'IA est à l'origine d'une bonne partie de nos fantasmes et peurs sur l'IA. Elle est justifiée dans la mesure où une bonne partie du savoir exploité par l'IA est d'origine humaine. C'est en limitant cet antropomorphisme à la fois dans nos projections et dans la création de systèmes à base d'IA que l'on peut revenir sur un terrain de confiance vis-à-vis de cette dernière.

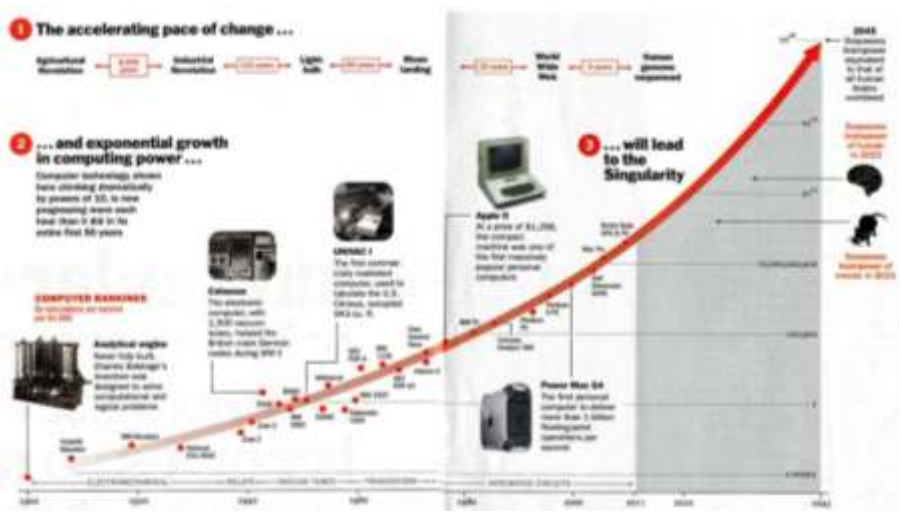
Les inquiétudes faceà l'IA sont peut-être aussi amplifiées par la peur de ne pas maîtriser la solution, de ne pas être en capacité de prévoir les résultats et de se sentir dépassé par une technologie qui dépasse les capacités du cerveau humain. Cela peut créer une sensation de perte de maîtrise qui accentue le sentiment de peur.

Risques

Avec l'IA, l'Homme a peur d'être dépassé par ses propres créations, peur de perdre le contrôle de son devenir, à la fois intellectuel et dans la maîtrise du monde physique. Pourtant, les évolutions technologiques passées n'avaient pas cet impact. L'Homme est d'ailleurs déjà largement dépassé par ses propres créations depuis longtemps[1738], d'abord du côté de la force physique, puis de calcul, de mémoire et, enfin, de traitement. Les machines mécaniques dépassent la puissance humaine depuis des lustres. On s'y est habitué et on s'en accomode. Ce sont des outils. Il n'y a que dans le raisonnement abstrait que les machines ont encore du mal à se développer.

Cette peur est alimentée par la perspective de voir émerger d'ici à peine quelques décennies une IA généraliste quelque peu mythique (AGI = artificial general intelligence, elle-même générant une ASI, Artificial Super Intelligence) omnisciente, omnipotente et contrôlant tout notre monde physique, et qui pourrait en retour nous asservir. Cette peur s'appuie sur une extrapolation des capacités actuelles de l'IA.

Elle s'appuie notamment sur une vision très simpliste d'une l'application ad-vitam de la loi de Moore (*ci-contre*).



L'intelligence humaine ne se réduit pas à un nombre de transistors ou de neurones.

Plus on rentre dans le détail du fonctionnement des neurones, plus on découvre leur complexité.

[1738] Et pas besoin d'en ajouter avec des annonces comme New AI Can Write and Rewrite Its Own Code to Increase Its Intelligence, février 2017,

Depuis les prédictions de Ray Kurzweil en 2005 sur la singularité, la complexité apparente des neurones a été multipliée par 100[1739].

Plus les futurologues de l'IA ont de véritables connaissances scientifiques en IA et en neurosciences, moins ils en ont peur. Nombre d'entre eux considèrent même que l'AGI, tout comme la singularité, sont des mythes.

Vous avez certainement entendu parler des craintes de l'IA manifestées par Bill Gates, Elon Musk, Stephen Hawking, Yuval Harari ou Henri Kissinger[1740]. Ils sont certes connus et respectés[1741]. Mais ils ont ceci de commun de ne pas être des spécialistes de l'IA[1742]. Yann Le Cun et Luc Julia sont eux beaucoup plus optimistes, ou moins pessimistes, comme vous le voulez[1743]! Jack Ma d'Alibaba est pour sa part plus optimiste[1744].

Dans des scénarios de prospective dignes des meilleures dystopies de science fiction, la première AGI générerait d'elle-même une ASI (Artificial Super Intelligence) qui prendrait le contrôle de la planète et annihilerait toutes les autres AGI, physiquement, via un contrôle direct des infrastructures, ou via divers « hacks ». Cette anticipation est une vue de l'esprit très centralisatrice[1745].

Certes, Google domine l'Internet occidental avec son moteur de recherche et Facebook domine les résaux sociaux occidentaux mais Internet reste assez distribué en l'état.

Comme avec l'usage de nombreuses autres technologies (automobile, armes à feu…), le vrai danger est l'Homme lui-même. C'est lui qui décide ce qu'il va en faire[1746]. C'est lui qui créé les règles utilisées par l'IA. Elles peuvent respecter les lois de la robotique d'Asimov… ou pas ! L'IA est plus ou moins dangereuse selon ce que l'Homme lui donne à faire et à décider[1747]. Si un chatbot s'entraîne en discutant avec des internautes racistes, il deviendra lui-même raciste comme l'a malheureusement expérimenté Microsoft avec Tay en 2016. Bref, l'enfer, c'est l'Homme, pas l'IA !

[1739] C'est un ordre de grandeur aussi approximatif que ses prévisions !

[1740] Voir AI could mean the end of human history, de Henry Kissinger, dans The Atlantic, juin 2018.

[1741] Exemple avec cette conférence « Où va nous mener d'intelligence artificielle » de Georges Mitaut, en décembre 2016, ingénieur en télécommunications retraité, qui ressasse les lieux communs habituels sur l'IA et la singularité sous forme d'une revue de presse bien slidisée. Au milieu de sa présentation et à propos de Google se trouve un logo de TensorFlow. Le conférencier indique alors que cela fait de l'IA mais qu'il n'a aucune idée de la manière dont cela fonctionne ! Ce n'est pas une exception ! La majorité des conférenciers sur l'IA font de la prospective alors qu'ils ne connaissent même pas les mécanismes actuels de l'IA et l'état de la recherche fondamentale. C'est consternant !

[1742] Mais parfois, ils connaissent l'IA et ne sont pas d'accord. C'est le cas de Sergei Brin. Voir Google Cofounder Sergey Brin warns of AI's dark side tandis qu'Eric Schmidt, l'ancien chairman de Google, trouve qu'Elon Musk est trop pessimiste. Voir Google billionaire Eric Schmidt: Elon Musk is 'exactly wrong' about A.I. because he 'doesn't understand', mai 2018 ainsi que Elon Musk is Wrong, AI won't kill us all de Toby Walsh, septembre 2017.

[1743] Voir l'interview Pour moi, l'IA n'existe pas de Luc Julia dans CBNews en juin 2018. Et pour Yann Le Cun : IA : « Les scénarios à la Terminator n'existent pas », estime Yann Le Cun dans Numerama, octobre 2017, Yann Le Cun, de Facebook : « l'intelligence artificielle va sauver des vies », dans Le Monde, septembre 2017 puis Yann Le Cun : "L'intelligence artificielle a moins de sens commun qu'un rat", dans Sciences et Avenir, janvier 2018.

[1744] Voir Jack Ma and Elon Musk hold debate in Shanghai, août 2019 (46 minutes).

[1745] Voir cette intéressante analyse des biais cognitifs de ces prévisions alarmistes : The Seven Deadly Sins of Predicting the Future of AI de Rodney Brooks, septembre 2017.

[1746] Voir Will AI kill us all after taking our jobs? de Fabio Ciucci publié en juillet 2017 ainsi que AI (Deep Learning) explained simply publié en juin 2017.

[1747] Dans « Life 3.0 », Max Tegmark évoque la thèse de Nick Bostrom dans « Superintelligence » selon laquelle le but ultime d'un système est indépendant de son intelligence, qui fournit les moyens d'atteindre un objectif complexe indépendamment de sa nature. L'intelligence humaine est ainsi déjà mise au service du bien commun et de la science ou du mal sous toutes les formes. Aux USA, ce dernier cas relève des « smart assholes » ! A la fin de son ouvrage, Max Tegmark indique que pour garder le contrôle de l'IA, il faut qu'elle apprenne, adopte puis conserve les objectifs qui lui ont été assignés par les Humains.. Voir Life 3.0 by Max Tegmark review – we are ignoring the AI apocalypse, septembre 2017, une review du livre de max Tegmark par Yuval Harari.

L'IA présente des risques bien plus prosaïques, comme toutes les technologies numériques : dans sa sécurité et dans ce qu'on lui demande de faire ! La sécurité d'un système d'IA peut être compromise à plusieurs niveaux : dans les réseaux et le cloud, dans les capteurs, dans l'alimentation en énergie.

Les bases de connaissances peuvent aussi être induites en erreur par l'injection d'informations erronées dans des images qui visent à altérer son comportement, par exemple dans le cadre d'un diagnostic médical complexe. On peut imaginer l'apparition d'anti-virus spécialisés pour les logiciels de deep learning.

Comme le code et les logiciels régissent de plus en plus notre vie, ce sont les règles qu'ils exécutent qui organisent la société. L'IA est aussi basée sur des règles (pour les moteurs de règles et les réseaux d'agents) et sur le mimétisme des sens et comportements humains (pour le deep learning).

Les dangers potentiels d'une hypothétique AGI sont surtout liés aux interactions entre les machines l'exécutant et le monde extérieur. Un robot logiciel n'est pas dangereux s'il tourne dans une machine isolée. Il peut le devenir s'il contrôle une arme de destruction massive ou des infrastructures stratégiques dans le monde extérieur et qu'il est programmé par des forces maléfiques ou simplement inconscientes.

La capacité à débrancher une AGI est devenue un thème de recherche des plus sérieux. Google annonçait en janvier 2016 qu'il travaillait sur la notion de « kill switch » de l'IA sans que l'on en connaisse d'ailleurs la nature. On sait juste que ce sont des chercheurs de DeepMind qui étudient des scénarios d'interaction entre robots et hommes dans des situations sous contrainte multiples : assurer une tâche d'un côté et réagir à des imprévus d'autre part.

Le "kill switch" de l'AGI qui permettrait de la déconnecter si elle devenait dangereuse devrait surtout porter sur sa relation avec le monde physique. Même si les films de science fiction tels que Transcendance rappellent que rien n'est sûr de ce côté là et que la tendance à tout automatiser peut fournir un trop grand contrôle du monde réel aux machines. Reste à trouver le « kill switch » pour une IA fortement distribuée, comme Skynet !

Je m'étonne toujours de notre capacité à construire des paquebots, porte-containers et porte-avions de 300 à 400 m de long et pesant plus de 100 000 tonnes. Un tableur compte plus vite que n'importe quel champion de calcul mental, ce depuis 1979. Autant la capacité de traitement parallèle d'un cerveau humain est impressionnante, autant sa capacité de stockage est limitée.

Une simple clé USB de 32 Go peut contenir plus de textes que ce que nous lisons, écrivons, entendons et disons pendant toute notre vie [1748]! Et ce que nous retenons ne fait qu'environ 1 Go ! Plus les outils numériques stockent l'information et sont faciles à utiliser pour l'interroger, moins on la retient. L'Homme continue de déléguer aux machines des tâches de plus en plus nombreuses, y compris celles qui le relient à l'espace pour l'appréhender. Ainsi, il aurait été démontré chez les chauffeurs de taxi utilisant Waze pour se déplacer que cela diminuait certaines zones du cerveau et entraînait des lacunes de mémoire. On le constate souvent avec les chauffeurs Uber. Cela a de nombreux effets pervers, comme cette histoire de touristes japonais qui ont laissé couler dans l'océan leur véhicule de location. Leur système de navigation avait été défaillant[1749].

[1748] Je me suis amusé à faire le calcul suivant : une personne qui vit 85 ans représentant 31 025 jours, pendant lesquels elle va lire 100 pages par jour, en écrire 20 par jour, et parler ou écouter parler pendant 8 heures par jour à raison de 200 mots à la minute va générer ou être exposée à 41 Go de données textuelles. C'est évidemment un cas extrême. Pour les gens moins bavards, moins lecteurs et moins producteurs, cela va tomber largement à moins de 10 Go. Ce qui ne fait pas grand-chose ! Qui plus est, on ne retient qu'une toute petite portion de tout cela. Donc, notre mémoire verbale ne fait probablement qu'à peine qu'environ 1 Go, avec cependant de nombreuses ramifications, les « hyperliens » du cerveau. Cette évaluation à la louche n'intègre pas la mémoire visuelle et auditive, qui est bien plus dense. La mémoire est notamment limitée par la durée de notre vie et la vitesse d'accumulation de connaissances et expériences.

[1749] Voir L'IA doit être contrôlée, mais les humains sont parfois faignants par Daphne Leprince-Ringuet, février 2020.

L'IA est aussi anxiogène car elle peut générer des systèmes pérennes dans le temps. Ses processus d'apprentissage bénéficient de la mémoire presque infinie des machines. L'IA serait donc immortelle, tant que ses systèmes de stockage ne défaillent pas.

On peut se rassurer en rappelant qu'un disque dur peut planter à tout bout de champ au bout de cinq ans et qu'un disque SSD actuel ne supporte au mieux que 3000 cycles d'écriture ! Mais leur remplacement robotisé est tout à fait possible dans des datacenters.

Enfin, les data centers ont besoin d'énergie et ils sont encore rares à être autonomes de ce point de vue-là[1750].

Mais les forces obscures humaines veillent au grain. Quelle sera l'arme de destruction massive à base d'IA ?

L'autre peur, plus court terme, et que nous étudierons plus loin concerne les évolutions des métiers qui, soit disparaîtront, soit deviendront bien plus productifs grâce à l'IA. C'est une crainte économique et sociale plus qu'une crainte de perte de contrôle de l'IA par l'Homme.

En attendant cette hypothétique AGI, les plus grands risques d'usage de l'IA ont trait à la mauvaise utilisation des IA étroites actuelles, liée entre autres aux lacunes de compétences de leurs créateurs, surtout lorsque les Humains ne sont plus dans la boucle, ainsi que tous les mécanismes qui ont tendance à réduire la portée de l'intelligence humaine. Sans compter évidemment les biais des données.

Pessimistes

Il existe deux types de pessimistes sur l'impact de l'IA sur la société. On a d'un côté ceux qui s'inquiètent avant tout d'une IA mythique qui n'existe pas encore, l'AGI et les thèses associées de la singularité, et puis ceux qui craignent l'IA d'aujourd'hui avec ses biais, son inexplicabilité et les risques immédiats qu'elle fait peser sur l'emploi et sur les démocraties.

En 340 avant JC, le philosophe grec **Aristote** prédisait déjà que des machines intelligentes capteraient le travail des humains. Et l'armateur britannique **Richard Thornton** anticipait en 1847 que les machines contrôleraient le monde, une prévision réitérée par **Alan Turing** en 1950.

De manière contemporaine, la principale source de pessimisme concernant l'impact de l'IA sur l'humanité vient cependant de la science fiction. Si l'on observe la production cinématographique des dernières décennies, les dystopies prennent largement le dessus des utopies. L'utopique Bicentenial Man en 1999 a été un flop tandis que tous les Terminator et son Skynet ont été des blockbusters[1751].

Ceci étant dit, les films racontant la vie heureuse de familles de cadres aisés avec trois enfants sont moins fréquents que les films d'horreur ou les policiers en tout genre. Les morts parmis les agents de la CIA dans certains films et certaines séries d'espionnage (Jason Bourne, 24, Scandal, etc) permettraient de remplir plusieurs murs de mémorial de la CIA !

[1750] Un data center alimenté par sa propre centrale nucléaire serait très dangereux. Et il n'existe pas de data centers alimentés entièrement par des panneaux solaires photoltaïques. Leurs onduleurs permettent en général de tenir quelques heures ou journées sans alimentation électrique.

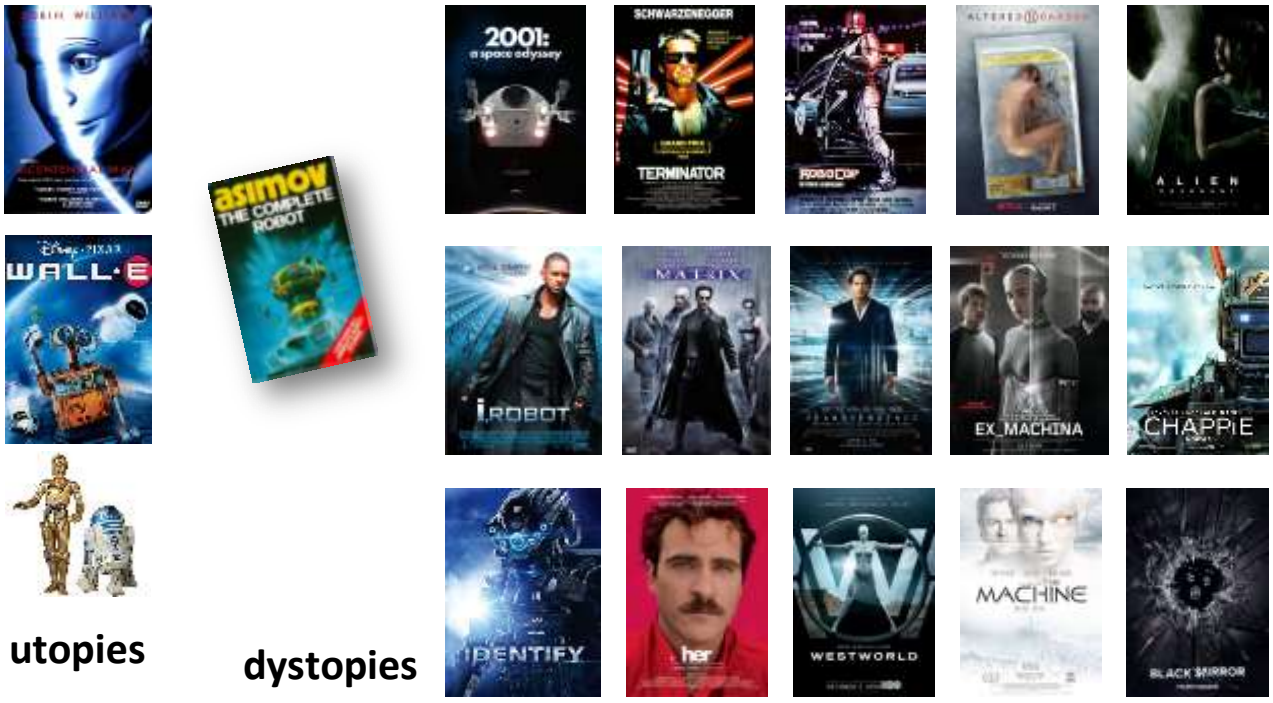
[1751] Même si on passe de la dystopie à l'utopie pour ce qui est du rôle de ce robot à partir du second film. Avec un bon robot contre un mauvais robot, la dualité bien/mal humaine reproduite dans les machines !

Et quand la science se fait fiction, elle déraille presque systématiquement et l'Homme en perd le contrôle, comme dans la série Westworld[1752].

Malgré tout, les grands auteurs de science fiction tels qu'Isaac Asimov, ont été prescients dans un grand nombre de leurs fictions. Ils l'ont d'ailleurs été plus sur l'impact sociétal des technologies de la communication que sur les technologies avancées dans d'autres domaines (transports, santé, …). Ce sont souvent les technologies qui dépendent d'infrastructures lourdes qui ne sont pas déployées.

Les alertes sur les risques de l'IA gagnent en écho lorsqu'elles proviennent de personnalités scientifiques et entrepreneuriales.

L'astrophysicien **Stephen Hawking** n'hésitait pas à prophétiser en 2014 que lorsque l'IA dépassera l'intelligence humaine, ce sera la dernière invention humaine, celle-ci ayant ensuite pris entièrement le pas sur l'espèce humaine [1753]!



Il reprenait à son compte une citation d'**Irwin John Good** de 1965[1754] selon laquelle la machine ultra-intelligente sera la dernière invention que l'homme aura besoin de créer (*ci-dessous*). Pour autant, si Hawking s'y connaît bien en trous noirs, il n'est pas forcément spécialisé en réseaux de neurones et deep learning.

9. Conclusions

These "conclusions" are primarily the opinions of the writer, as they must be in a paper on ultraintelligent machines written at the present time. *In the writer's opinion then*:

It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make, since it will lead to an "intelligence explosion." This will transform society in an unimaginable way. The first ultraintelligent machine will need to be ultraparallel, and is likely *to be achieved with the help of a very large artificial neural net.*

Cette thèse se retrouve décrite dans le menu dans de nombreux ouvrages, comme **Superintelligence** de Nick Bostrom paru en 2014 ou dans **Our Final Invention, Artificial Intelligence and the End of the Human Era** de James Barrat, paru en 2015.

[1752] Voir Westworld, as reviewed by scientists, roboticists, researchers par Jennifer Bisset, juillet 2018, et aussi The incredible, unbelievable, rapidly advancing future, and the end of The End de Ben Edwards, octobre 2017, définit quelques-uns des scénarios catastrophes d'une IA dont l'Homme perdrait le contrôle.

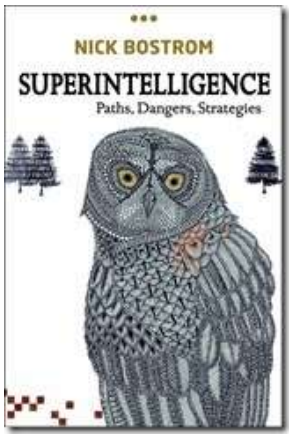
[1753] Voir Stephen Hawking warns artificial intelligence could end mankind, mars 2014.

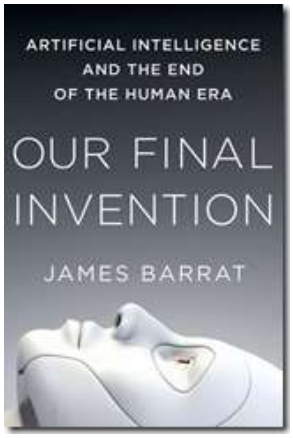
[1754] Publiée dans Speculations Concerning the First Ultraintelligent Machine, 1965 (30 pages).

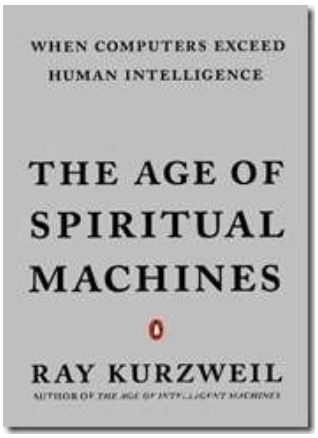
Ces prédictions partent du principe que l'on arrivera un jour à créer une machine superintelligente dont la puissance croitra de manière exponentielle et qui contrôlera toutes nos destinées du fait de l'hyperconnexion des infrastructures physiques et des objets de la vie courante.

Le cofondateur de Sun Microsystems, **Bill Joy**, avait été l'un des premiers à alerter l'opinion en 2000[1755], sur les dangers des progrès technologiques dans l'IA, les nanotechnologies et les biotechnologies.

C'était bien avant la fin du premier séquençage complet du génome humain qui avait coûté une fortune[1756]. Bill Joy était en fait effrayé des perspectives avancées par Ray Kurzweil qu'il avait rencontré dans une conférence en 1998 et après avoir lu son **The age of spiritual machines**, paru six ans avant **The singularity is near**.







S'en est suivie une grosse décennie de calme côté alertes. Après Stephen Hawking en 2014, Bill Gates et Elon Musk ont repris le flambeau de Bill Joy en 2015 pour demander une pause technologique et une réflexion sur les limites à ne pas dépasser avec l'intelligence artificielle comme avec la robotique. Pause de quoi précisément ? Ce n'était pas bien clair. Peut-être pour les rares startups bien financées qui planchent sur l'AGI comme Numenta. Mais aucune pause n'a eu lieu. Par contre, cela a lancé le débat sur l'explicabilité de l'IA et poussé les gouvernements et leaders d'opinions à s'intéresser à l'éthique de l'IA.




[1755] Dans Why the future doesn't need us, Bill Joy, Wired, 2000.

[1756] On y apprend d'ailleurs qu'il avait rencontré Jacques Attali et que ce dernier avait indirectement influé le cours des événements de Java !

Il existe même des instituts de recherche qui planchent sur la question des risques de l'IA, tels le **Center for the Study of Existential Risk** de Cambridge et le **Future of Humanity Institute** d'Oxford.

Les dangers perceptibles de l'IA sont à l'origine de la création d'**OpenAI**, fin 2015, une initiative visant non pas à créer une IA open source – les grands logiciels de l'IA sont déjà presque tous open source - mais à surveiller et analyser ses évolutions. Il s'agit d'une ONG créée par Elon Musk. Elle était dotée au départ de $1B de financement. Son objectif : faire de la recherche et de la formation. Dans certains cas, leurs actions semblent un peu éloignées de l'objectif initial, comme dans ce benchmark de jeux vidéo opposant des joueurs humains et artificiels. Pour Elon Musk, OpenAI devait s'assurer de manière assez manichéenne que l'IA fasse le bien et pas le mal, dans une vision assez naïve du fonctionnement du capitalisme. OpenAI milite en tout cas pour une régulation de l'IA[1757].

Microsoft annonçait en juin 2019 investir $1B dans OpenAI pour les aider à créer une AGI (intelligence artificielle générale). Une AGI for good comme il se doit ! Fin 2018, OpenAI changeait d'ailleurs de statut pour passer d'une entité à but non lucratif à un statut classique d'entreprise (Limited Partnership), lui permettant ainsi de récupérer ce genre de financement de la part de Microsoft. Elon Musk avait d'ailleurs quitté OpenAI en février 2019, à la fois en raison de désaccords non précisés avec la stratégie d'OpenAI et du fait de collisions possibles avec les travaux de Tesla. Elon Musk continue de faire la science fiction en dénonçant les risques de voir l'IA créer un dictateur immortel[1758] !

D'autres initiatives semblables ont vu le jour comme **Partnership on AI**[1759], une association créée en 2016 et rassemblant comme membres fondateurs tous les GAFAMI (Google, Facebook, Amazon, Microsoft, IBM, et Apple qui les a rejoints en janvier 2017). L'association est présidée par Eric Horvitz, le patron de Microsoft Research et Mustafa Suleyman de Google, le co-fondateur de DeepMind. Elle débat des questions soulevées par l'IA et des meilleures pratiques à adopter pour en mitiger les risques.

Quand on y regarde de près, cette association prend la forme d'une organisation de lobbying avec les méthodes associées : des thématiques à défendre, l'appel à des experts divers, l'organisation de débats et un pied dans la porte des politiques pour éviter des dérives réglementaires gênant l'innovation. Eric Horvitz promeut de son côté l'utilisation de l'IA pour le bien de l'humanité[1760].

[1757] Voir OpenAI Director Shivon Zilis explains why AI requires oversight now, novembre 2017.

[1758] Voir Elon Musk warns A.I. could create an 'immortal dictator from which we can never escape' par Ryan Browne, 2018.

[1759] Voir https://www.partnershiponai.org/.

[1760] Voir son support de présentation, bien documenté d'études de cas d'usages positifs de l'IA : AI in Support of People and Society, juin 2016 (81 slides).

Il faut aussi évoquer le méta-pessimisme de l'IA : c'est-à-dire le pessimisme lié à l'impact du pessimisme sur l'IA sur son développement. En effet, pour ceux qui considèrent que les peurs sur l'IA sont exagérées, celles-ci ont pour résultat de ralentir l'innovation et en particulier de mettre les pays occidentaux en position de faiblesse face à la Chine qui se pose moins de questions sur l'impact sociétal de l'IA. C'est une opposition courante face aux promoteurs d'une IA plus éthique[1761].

AINow est une autre initiative lancée mi 2016. Elle associe Kate Crawford (Microsoft Research) et Meredith Walker (Google Open Research Group).

Basée à New York, elle est est focalisée sur l'impact de l'IA sur les droits civiques, sur l'emploi, les biais et la sécurité des infrastructures. Elle a publié un premier rapport en 2016[1762].

Sous couvert de bonne gestion du principe de précaution, ces initiatives des GAFAMI sont à analyser sous la loupe des pratiques habituelles du lobbying[1763]. Elles visent aussi à calmer les peurs et à assurer les pouvoirs publics qu'une autorégulation de l'IA est possible par les acteurs de l'industrie.

Cela vise surtout à éviter que ces derniers s'immiscent dans la stratégie de ces grands acteurs. Et dans le cas où il viendrait à l'idée des pouvoirs publics de réguler l'IA d'une manière ou d'une autre, d'être prêt avec des propositions compatibles avec leurs stratégies. C'est de bonne guerre mais il ne faut pas être dupe !

Enfin, Elon Musk a aussi lancé fin 2016 sa startup **Neuralink** dont l'objectif est de relier l'IA à l'homme pour éviter d'en perdre le contrôle[1764], via des nano-électrodes directement implantées dans le cerveau et capables d'activer sélectivement les neurones. Ce projet servira probablement surtout à améliorer l'état de l'art du traitement de certaines pathologies neurodégénératives diverses, qui ne nécessitent pas d'agir au niveau de neurones individuels[1765].

L'autre personnalité inquiète du futur de l'IA est le chercheur anglais **Stuart Russell**. Il veut faire réguler l'usage d'armes robots. Il a fait produire la vidéo de fiction glaçante Slaughterbots sur un hypothétique futur où des drones armés d'explosifs ciblent des dissidents dans un pays occidental. Elle était diffusée à l'occasion d'une conférence de l'ONU sur le sujet. Stuart Russell veut interdire ce genre de drones et toute forme de robot tueur[1766]. Le pire est que la technologie présentée dans la vidéo ne relève pas de la science fiction et est disponible dès aujourd'hui. Et d'ailleurs, l'US Air Force étudie déjà la question des drones autonomes[1767]. Et la branche de l'ONU en charge du désarmement s'intéresse à l'usage de l'IA dans les armes.

[1761] Voir AI's current hype and hysteria could set the technology back by decades par Wim Naudé, juillet 2019.

[1762] Voir The AI Now Report - The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term, A summary of the AI Now public symposium, hosted by the White House and New York University's Information Law Institute, juillet 2016 (37 pages) et Why AI is still waiting for its ethics transplant, de Scott Rosenberg dans Wired, novembre 2017.

[1763] Exemple récent : le lancement du challenge AI Impact doté de $25M pour récompenser des projets qui font de l'IA for good. Voir Google pledges $25 million toward AI solutions for social issues, octobre 2018.

[1764] L'idée est inspirée des neural laces de l'auteur de science fiction Iain M. Banks. Voir The novelist who inspired Elson Musk If you want to understand where society is heading, read the novels of Iain M. Banks, Silicon Valley's favourite author, de Tim Cross, mars 2017.

[1765] On peut aussi imaginer des solutions visant à activer les neurones de l'hippocampe qui est une sorte de gatekeeper de la mémoire. C'est lui qui transfère la mémoire court terme au sein du cerveau limbique vers la mémoire long terme du cortex en périphérie du cerveau.

[1766] Voir UN to host first talks on use of 'killer robots', novembre 2017.

[1767] Voir The Air Force is exploring AI-powered autonomous drones, par Amrita Khalid, mars 2019.

Dans le top du top des prévisions, il y a aussi celle-ci qui prévoit qu'eu 2042 une IA va créer son dieu et sa propre bible. Elle vient d'un certain **Anthony Levandowski** qui a déposé les statuts d'une organisation religieuse "The Way of the Future » qui ambitionne de créer un dieu basé sur une IA qui ferait le bien de la société[1768]. 2042, ça doit être une blague de Douglas Adams ! Si cela se trouve, le risque de l'IA sera plus basique. L'abus d'IA pourrait nous enfermer dans notre passé et limiter la sérendipité de notre vie qui deviendrait toute lisse.

On continuerait à se fier de plus en plus aux machines pour organiser notre vie comme on le fait déjà avec les outils de communication qui nous interrompent sans cesse. Cet abandon progressif de notre libre arbitre serait consenti volontairement et progressivement. L'IA abuserait de nos faiblesses pour nous orienter dans nos choix comme le fait déjà la publicité qui exploite sans vergogne nos pulsions[1769].

la vie des utilisateurs est déjà largement régie par les logiciels et les plateformes

Pour ce qui maintenant est des menaces à plus court terme de l'IA, il existe même une expression : la **dark AI**, ou « IA sombre » voire « IA obscure » pour suivre la nomenclature StarWarzienne. Elle est alimentée par les risques de création de nouveaux monopoles ou d'amplification de monopoles exisants, de manipulation des consommateurs et électeurs, d'augmentation de la fracture numérique et de viol des vies privées. Cela a même engendré la création d'initiatives internationales diverses : à l'**ONU** (Centre for AI and Robotics de l'United Nations Interregional Crime and Justice Research Institute, UNICRI), au **WEF** (World Economic Forum), G20 et à l'**OCDE**[1770].

Le philosophe **Miguel Benasayag** parle de tyrannie des algorithmes. Mais surtout de manière méta, en évoquant la vision « machinique » de la société qu'elle engendre, réduisant les individus à des « avatars virtuels » alors qu'une vie est bien plus aléatoire que les modèles créés à partir de procédés répétitifs[1771]. Tout cela fait dire à **Jaron Lanier** que l'IA est plus une idéologie qu'une technologie[1772]. Seulement voilà, c'est une idéologie observée mais pas construite explicitement, même si certains tentent d'expliquer politiquement et idéologiquement la structure entrepreneuriale de la Silicon Valley. Sachant que cette structure est ensuite répliquée presque à l'identique dans le monde entier.

[1768] Voir An AI god will emerge by 2042 and write its own bible, will you worshipt it?, octobre 2017.

[1769] Voir Les péchés capitaux dans le marketing et la vente, Olivier Ezratty, 2011, Comment l'intelligence artificielle va rendre notre vie terriblement ennuyeuse, de Céline Deluzarche, juin 2018 ainsi que Evoque Re-Engineering Humanity de Brett Frischmann et Evan Selinger, avril 2018.

[1770] Voir How To Combat The Dark Side Of AI par Mark Minevich, février 2020.

[1771] Voir La Tyrannie des algorithmes par Miguel Benasayag, octobre 2019.

[1772] Voir AI is An Ideology, Not A Technology par Jaron Lanier dans Wired, mars 2020.

Ces menaces me semblent toujours pointer trop du doigt l'IA comme un concept générique. Le plus souvent, ce qui est à redouter concerne les implications des outils numériques en général, les briques d'IA n'en étant que certains des aspects les plus saillants. Les biais des algorithmes ? Il y en a toujours eu ! Les menaces sur la vie privée ? On en parlait déjà il y a plusieurs décennies. En voici une illustration avec un article du Nouvel Observateur de 1969 qui titrait « *La fin de la vie privée* » au sujet des méga-ordinateurs de l'époque, notamment les mainframes d'IBM de la série S/360 et dont les capacités de stockages ne se comptaient pas encore en Go. Cela ne veut pas dire que les menaces n'existent pas mais que l'on a parfois tendance à crier au loup un peu trop tôt. On a tout de même eu la loi « Informatique et libertés » votées en 1978 en France, à l'origine de la création de la CNIL.


La fin
de la vie privée

« Le méga-ordinateur est pour demain : avec lui, il n'y aura plus aucune chance, non seulement d'avoir une vie privée, mais d'avoir une vie tout court

Optimistes

Les optimistes semblent moins nombreux. On y trouve bien évidemment les singularistes dont le pape actuel, **Ray Kurzweil**, anticipe l'émergence d'une AGI autour de 2030-2040 en nous promettant monts et merveilles qui peuvent nous encourager à procrastiner sur la résolution des problèmes d'aujourd'hui (réchauffement climatique, surpopulation, inégalités, …).

La procroastination a d'ailleurs lieu sans qu'il en soit responsable. Donald Trump n'a certainement pas lu Kurzweil[1773] !

Pour le sceptique éclairé **Piero Scaruffi**, nous sommes tiraillés entre deux extrêmes de science-fiction avec des pessimistes qui pensent que l'IA va tous nous tuer et des singularistes qui estiment qu'elle va nous rendre immortels.

Au milieu de l'échiquier des optimistes se situent des personnalités telles que **Mark Zuckerberg** qui estime que l'Homme sera raisonnable dans ses usages de l'IA[1774] et puis **Ginni Rometti** d'IBM qui recommande de ne pas avoir peur des robots[1775]. D'autres comme **Sarah Kessler** prévoient que la transformation des métiers générera un nouvel équilibre, pas forcément moins bon que l'actuel, et qu'il n'y a pas lieu de s'inquiéter[1776]. Enfin l'auteur **Sam Harris** pense que l'on peut très bien conserver le contrôle de l'IA[1777].

Les optimistes sont aussi souvent les véritables spécialistes de l'IA qui voient de près l'ingratitude de la discipline et estiment en général que l'on est très éloigné de l'AGI et de l'ASI.

[1773] Ray Kurzweil est présenté selon les circonstances comme le directeur de la recherche de Google, son patron de la R&D en IA, directeur de l'engineering quand ce n'est pas « chief futurist ». Alors que les deux principales équipes d'IA de Google, Google Brain chez Google X et celle de DeepMind ne dépendent pas de Kurzweil. Il n'a rien produit ou annoncé depuis son arrivée chez Google en 2012. On sait qu'il planche sur le traitement du langage avec une équipe d'une vingtaine de personnes, à comparer aux 300 personnes qui travaillent chez DeepMind. Il travaillerait à la création d'un chatbot qui répondrait automatiquement à nos emails à notre place. Voir What is Ray Kurzweil up to at Google ? Writing your emails, février 2017. Il a fait une apparition au CEBIT 2017 où il a passé le plus clair de son temps à rappeler les effets de la loi de Moore et de ses déclinaisons dans d'autres domaines. Et d'évoquer quelques avancées dans la compréhension du fonctionnement du cerveau. On a surtout pu remarquer qu'il a maintenant une chevelure de quadra alors qu'il était quasiment chauve avant. Grâce aux plus de cent pillules qu'il prend chaque jour depuis des années pour prolonger sa durée de vie, à une greffe ou à une perruque ? Ray Kurzweil a aussi piloté le projet Talk to Books, permettant de poser des questions « à un livre » qui va y trouver les réponses … ou pas.

[1774] Voir Intelligence artificielle : Zuckerberg (Facebook) n'a pas peur, dans ZDNet, février 2016.

[1775] Voir At Davos, IBM CEO Ginni Rometty Downplays Fears of a Robot Takeover de Claire Zillman dans Fortune, janvier 2017.

[1776] Voir The optimist's guide to the robot apocalypse de Sarah Kessler, mars 2017.

[1777] Voir la video TEDx Can we build AI without losing control over it par Sam Harris (14 mn).

La plupart des auteurs qui prédisent une ASI ne sont en effet pas des spécialistes de l'IA[1778] ! Les entrepreneurs de l'IA sont d'ailleurs fort agacés des prévisions délirantes concernant l'IA[1779].

Le principal moyen de s'en éloigner est de faire la distinction entre l'association de l'intelligence humaine avec sa chair et ses sens, et n'importe quelle forme d'intelligence intégrée dans une machine dénouée de cette chair et de ces sens.

L'un des écueils principaux des prévisions pessimistes est leur anthropomorphisme[1780], tourné dans le mauvais sens [1781]!

Le deep learning exploite souvent de l'expertise d'origine humaine, dans DeepMind AlphaGo aussi bien que dans les systèmes de reconnaissance d'image en imagerie médicale, Il en va de même pour IBM Watson en cancérologie qui exploite la littérature scientifique d'origine humaine sur le sujet. L'IA applique une force brute sur une vaste base de données d'intelligence humaine.

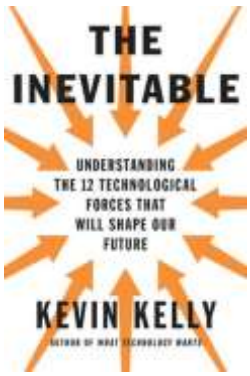
L'apprentissage supervisé fonctionne par imitation et utilise des tags d'origine humaine. Et l'apprentissage non supervisé, comme dans les premières phases des réseaux de neurones convolutifs exploite en bout de course de l'apprentissage supervisé. Idem pour les modèles génératifs qui appliquent des styles d'origine humaine à divers contenus.

La soi-disante créativité des réseaux de neurones génératifs s'appuie toujours sur la créativité d'origine humaine qu'elle ne fait que répliquer machinalement et de manière prédictible ! Bref, en matière d'intelligence, l'IA imite le plus souvent des composantes très isolées de celle de l'homme qu'elle met en forme, en la démultipliant par la force brute.

Pour s'écarter de cet antropomorphisme, on peut adopter la posture de **Kevin Kelly**, auteur du best seller « The inevitable », et qui considère que l'IA doit être considérée comme « alien »[1782]. On peut aussi écouter les envolées lyriques du philosophe **Grady Booch** qui explique pourquoi il ne faut pas avoir peur de l'IA (vidéo), que l'Homme entraînera à ne pas lui nuire.

conférence sur les usages "for good" de l'IA


Grady Booch



[1778] La situation s'inverse entre Elon Musk (pessimiste) et Mark Zuckerberg (optimiste) lorsque le premier accuse le seccond d'ignorance sur l'IA. Et on revient au point initial lorsque le roboticien Rodney Brooks contredit Elon Musk. Dans This famous roboticist doesn't think Elon Musk understands AI, TechCrunch, juillet 2017.

[1779] Voir par exemple La supériorité de l'intelligence artificielle : l'arnaque du siècle, de Denis Fagès, fondateur de VideoTelling, juillet 2018. Et aussi We're told to fear robots. But why do we think they'll turn on us? The robot uprising is a myth de Steven Pinker, février 2018 ainsi que Beware the AI delusion de Gary Smith, octobre 2018.

[1780] Comme cette petite tribune de Stéphane Mallard, Digital Evangelist, comme quoi l'IA sera capable de tout faire et d'être créative. L'IA, plus créative que l'Homme ?, dans l'ADN, avril 2017. Et on se fait avoir régulièrement par l'IA ! Voir 'Artificial Intelligence' Has Become Meaningless de Ian Bogost dans The Atlantic, mars 2017.

[1781] Encore une savoureuse citation de Piero Scaruffi : "*In private conversations about "machine intelligence" i like to quip that it is not intelligent to talk about intelligent machines: whatever they do is not what we do, and, therefore, is neither "intelligent" nor "stupid" (attributes invented to define human behavior). Talking about the intelligence of a machine is like talking about the leaves of a person: trees have leaves, people don't. "Intelligence" and "stupidity" are not properties of machines: they are properties of humans. Machines don't think, they do something else. Machine intelligence is as much an oxymoron as human furniture. Machines have a life of their own, but that "life" is not human life.*".

[1782] Voir Le mythe de l'IA surhumaine de Rémi Sussan, mai 2017. Kevin Kelly décrit cela lui-même dans Wired en avril 2017 : The myth of a superhuman AI. Principal argument : l'intelligence n'est pas unidimentionnelle.

De son côté, le chercheur français **Jean-Gabriel Ganascia**, auteur du « Mythe de la singularité » (2017) dénonce avec justesse la construction de mythes autour de l'IA et de la singularité[1783].

On peut aussi s'amuser de la crédulité de ceux qui ont avalé la création de **Rocket AI** (2016), une startup développant un réseau de neurones rappelant ceux de Numenta et baptisé « Temporal Recurrent Optimal Learning » (TROL). Il s'agissait d'une grosse blague de potaches de l'IA[1784] soulignant la crédulité de l'écosystème de l'innovation.

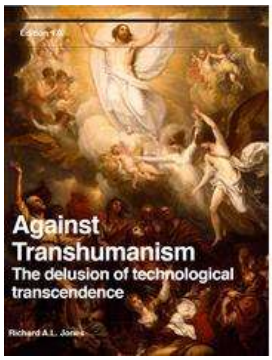
Autre méthode, se rassurer une fois encore avec les écrits de **Piero Scaruffi**[1785]. Ce dernier cherche à démontrer que la singularité n'est pas pour demain. Il s'appuie pour cela sur une vision historique critique des évolutions de l'intelligence artificielle.

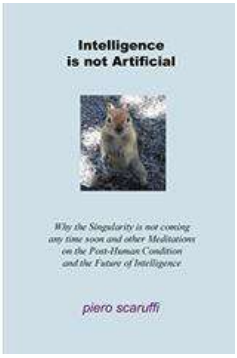
Il pense que les progrès de l'IA proviennent surtout de l'augmentation de la puissance des machines, et bien peu des algorithmes (ce qui serait à nuancer…). Il relativise les performances actuelles de l'IA, montées en épingle par les entreprises, les experts et les médias.

Selon lui, l'Homme a toujours cherché une source d'intelligence supérieure, qu'il s'agisse de Dieux multiples ou unique, de Saints ou d'extra-terrestres. La singularité et les fantasmes autour de l'IA seraient l'incarnation d'une nouvelle forme de croyance voire même de religion, une thèse aussi partagée par Jaron Lanier, un auteur anticonformiste qui publiait **Singularity is a religion just for digital geeks** en 2010[1786].

Piero Scaruffi prend aussi la singularité à l'envers en avançant que l'ordinateur pourra fort bien dépasser l'Homme côté intelligence parce que les technologies rendent Homo Sapiens plus bête[1787], en le déchargeant de plus en plus de fonctions intellectuelles, la mémoire en premier et le raisonnement en second !

Selon lui, le fait que les médias numériques entraînent les jeunes à lire de moins en moins de textes longs réduirait leur capacité à raisonner. A tel point qu'il devient impossible d'expliquer les effets de la baisse d'attention du fait de cette dernière[1788] !





[1783] Voir Technologie : peut-on se défaire des promesses et des mythes ?, une excellente revue de lecture de l'ouvrage de Jean-Gabriel Ganascia ainsi que de l'ouvrage collectif « Pourquoi tant de promesses » dirigé par Marc Audétat, par Hubert Guillaud, juin 2017.

[1784] Voir Rocket AI: 2016's Most Notorious AI Launch and the Problem with AI Hype, décembre 2016. Le site de Rocket AI n'est d'ailleurs pas moins documenté que celui de nombreuses startups de l'IA.

[1785] Comme Demystifying Machine Intelligence.

[1786] Voir Singularity Is a Religion Just for Digital Geeks, 2011.

[1787] Thèse partagée par **Daniel C. Dennett**, pour qui le véritable danger n'est pas dans les machines plus intelligentes que l'homme mais plutôt dans le laisser-aller de ce dernier qui abandonne son libre arbitre et confie trop de compétences et d'autorité à des machines qui ne lui sont pas supérieures.

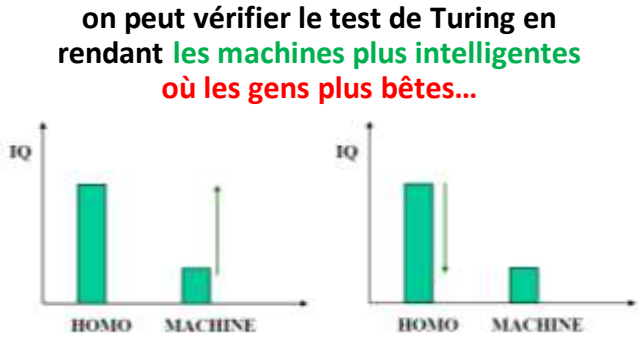
[1788] "*I am worried that people's attention span is becoming so short that it will soon be impossible to explain the consequences of a short attention span. I don't see an acceleration in what machines can do, but i do see a deceleration in human attention… if not in human intelligence in general*", dans "Intelligence is not artificial".

On peut d'ailleurs le constater dans les débats politiques qui évitent la pensée complexe et privilégient les simplismes à outrance. J'aime bien cet adage selon lequel l'intelligence artificielle se définit comme étant le contraire de la bêtise naturelle. Cette dernière est souvent confondante et rend le défi de la création d'une intelligence artificielle pas si insurmontable que cela dans un bon nombre de domaines.

Pour Piero Scaruffi, en tout cas, l'intelligence artificielle est d'ailleurs une mauvaise expression. Il préfère évoquer la notion d'**intelligence non humaine**. Une bonne approche qui souligne la complémentarité de l'IA et des Hommes.

Il pense aussi qu'une autre forme d'intelligence artificielle pourrait émerger : celle d'hommes dont on aura modifié l'ADN pour rendre leur cerveau plus efficace. C'est un projet du monde réel, poursuivi en Chine où sont séquencés des milliers d'ADN humains pour identifier les gènes de l'intelligence ! Histoire de réaliser une (toute petite) partie des fantasmes délirants du film Lucy de Luc Besson !

L'intelligence humaine cumule la capacité à créer des théories expliquant le fonctionnement physique du monde et à mener des expériences permettant ensuite de les vérifier.



"***Know-nothingism*** *is the insistence that there are simple, brute-force, instant-gratification answers to every problem, and that there's something effeminate and weak about anyone who suggests otherwise*"
Paul Krugman, New York Times, Aug 2008

Parfois, cette vérification s'étale sur un demi-siècle à un siècle, comme pour les ondes gravitationnelles ou l'existence du boson de Higgs. Cette capacité de théorisation et d'expérimentation de long terme n'est pour l'instant pas accessible à une machine, quelle qu'elle soit. Les machines ne se posent pas encore de questions existentielles sur leur relation au monde qui les entoure. Elles n'ont pas de sens commun.

Emplois et IA

Prenons maintenant un peu de recul sur la robotisation en marche des métiers liée aux avancées de l'intelligence artificielle vues jusqu'à présent. Cette robotisation n'a pas besoin d'AGI ou de singularité pour se poursuivre. Et il faut distinguer la notion de robotisation qui sous-entend le remplacement des humains, de l'outillage, qui leur permet d'être plus efficace.

Nous allons ici évoquer les différentes études et prévisions sur le sujet, d'abord d'un point de vue quantitatif, puis qualitatif.

Prévisions

Les prévisions sur la destruction et la création d'emplois liées au déploiement de l'IA dans l'économie sont pléthoriques depuis 2013.

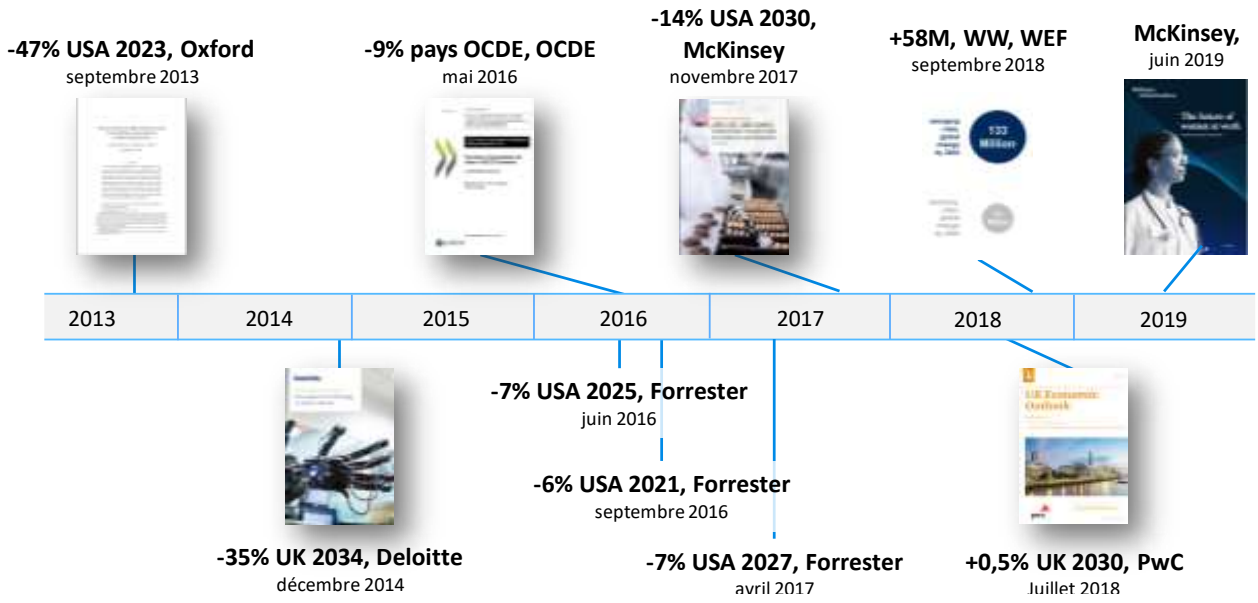
On y trouve aussi bien de sombres prophéties sur le rôle même de l'Homme dans l'économie que des prévisions plus optimistes, croyant fermement à la destruction-création de valeur schumpétérienne avec un équilibre positif.

La destruction nette d'emplois liée à l'IA à l'horizon 2023-2025 se situe selon les études entre 6% à 47%, avec des prévisions qui suivent une tendance baissière, la principale prévision de 47% datant de 2013 et celles de 6% à 7% datant de 2016. Ca donne une belle marge d'erreur et de manœuvre !

Cela illustre que les tendances lourdes sur le marché de l'emploi, si elles auront bien lieu, interviendront un peu plus tard. Pour que tel ou tel emploi disparaisse d'ici 5 ans, il faudrait que les technologies correspondantes soient disponibles aujourd'hui compte-tenu de l'inertie du marché, des parties prenantes, des budgets et des déploiements. Si elles ne sont pas encore disponibles, il faudra alors attendre bien plus de 5 ans pour qu'elles aient un impact sur l'emploi ! Un bon nombre de technologies mettent au moins entre 10 et 15 ans à se propager à l'échelle mondiale, surtout si elles nécessitent des infrastructures. Or nombre de prévisions s'appuient sur des technologies qui ne sont pas encore disponibles, même en amont de la R&D.

L'économiste **John Maynard Keynes** se faisait déjà l'écho des risques de pertes d'emploi liées à l'automatisation en 1933, avant même que les ordinateurs fassent leur apparition. Les premières prévisions sur les pertes d'emploi liées à l'IA sont arrivées dans les années 1960. Au démarrage des précédentes révolutions industrielles, les métiers disparus comme les nouveaux métiers ont rarement été bien anticipés.

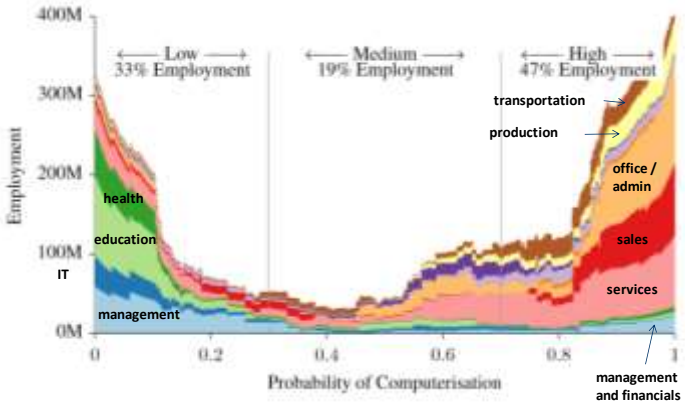
Pour ce qui est du futur, à vrai dire, on n'en sait pas grand chose. Les paramètres à prendre en compte sont tellement nombreux ! Malgré tout, on se doit de faire des prévisions, même fumeuses, pour anticiper la manière de préparer les générations futures.



La principale leçon à retenir des prévisions du passé est de conserver un peu d'humilité ! On peut cependant faire quelques hypothèses. Elles sont notamment utiles pour mener certaines politiques publiques, dans l'éducation comme dans les choix de développement infrastructures et de politique industrielle. On sait par exemple qu'il faudra privilégier la formation à des métiers qui ne sont pas répétitifs et où la créativité et l'adaptation jouent un rôle clé, sans compter l'empathie, le relationnel, le sens critique et le sens commun.

Je vais faire ici le tour des études qui font ces prévisions sur le futur de l'emploi et en décrire les méthodes et écueils le cas échéant.

- **Septembre 2013** : **Oxford** publie la première grande étude sur l'impact de l'IA sur le futur de l'emploi[1789], créée par Carl Benedikt Frey et Michael Osborne. Elle évoque la disparition de 47% des emplois aux USA à l'horizon de 2023[1790].



- **Août 2014** : **Pew Research Center** publie une étude qualitative qui recense l'avis de divers spécialistes dont certains estiment que la moitié des emplois sont menacés à l'horizon 2025[1791]. Ces experts sont très divisés sur la question[1792] !
- **Octobre 2015** : **Brookings Institution** et Darrell West prévoit des créations d'emplois dans plein de secteurs et des disparitions dans peu de secteurs[1793]. C'est une étude qualitative qui se penche sur les politiques publiques à appliquer. Il évoque le revenu minimum, des réformes fiscales diverses, des formations et le développement des activités culturelles !

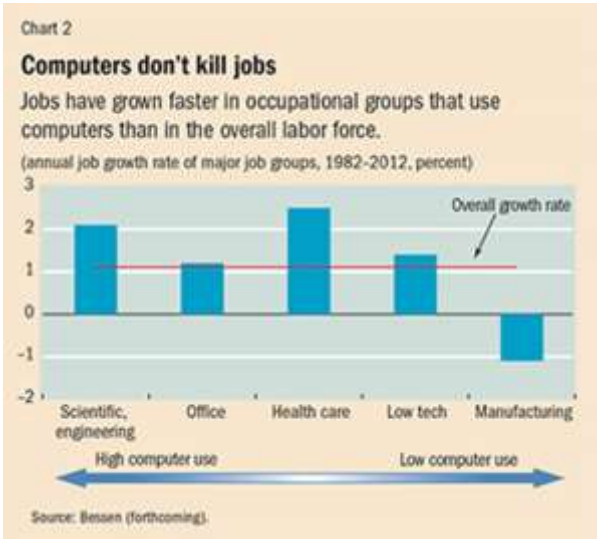
[1789] Voir The Future of Employment: How susceptible are jobs to computerisation?, septembre 2013 (72 pages).

[1790] L'étude segmente avec précision les métiers et leurs risques d'être remplacés par des machines. Le calcul du risque s'appuie sur trois formes d'intelligence clés des métiers : l'intelligence motrice (perception et manipulations), l'intelligence créative et l'intelligence sociale. On y constate que la situation est très polarisée : il y a d'un côté des métiers à très faible risque d'automatisation (<20%) comme les fonctions de management, dans la finance, dans le numérique, l'éducation et même la santé, et de l'autre, des métiers à très fort risque d'automatisation (>60%) et surtout dans les services, la vente et l'administratif. La méthode utilisée est mathématique et probabiliste. Nous sommes déjà cinq ans plus tard et les transformations en question sont loin d'avoir eu lieu. Les cycles de l'innovation sont bien plus lents que ce qu'ils prévoyaient. Leur étude sera peut-être valable mais à 15 ou 30 ans d'horizon. Cette étude a été souvent critiquée car elle raisonnait au niveau des métiers sans décomposer leurs tâches automatisables ou pas. Nombre d'études qui ont suivi ont repris cette méthodologie en la corrigeant.

[1791] Voir AI, Robotics and the future of jobs, août 2014 (67 pages).

[1792] Le pessimisme provient du risque d'impact rapide de l'automatisation sur les cols blancs avec un risque de déclassification pour un grand nombre, qui seront orientés vers des métiers moins bien payés. Enfin, le système d'éducation ne serait pas en mesure de s'adapter aux nouveaux enjeux. Certains experts sont optimistes car les métiers qui disparaissent sont naturellement remplacés par d'autres, au gré de l'évolution de la demande. La relation avec le travail sera aussi redéfinie de manière plus positive.

[1793] Voir What happens if robots take the jobs? The impact of emerging technologies on employment and public policy, de Darrell West, octobre 2015 (22 pages).



- **Mars 2015** : **FMI** et James Bessen décrivent l'absence de corrélation entre l'automatisation et les destructions d'emplois. C'est l'origine des schémas *ci-dessus* qui montreraient que les ordinateurs ne sont pas à l'origine de la suppression d'emplois[1794].
- **Décembre 2014** : **Deloitte** et **Oxford** publient une étude quantitative qui prévoit le remplacement de 35% des emplois au Royaume-Uni par des robots en 20 ans[1795]. Elle s'appuie sur l'Histoire et la destruction d'emplois dans le passé, notamment dans les métiers de secrétariat. Ils intègrent aussi les créations d'emploi et le solde a l'air d'être positif[1796].
- **Mai 2016** : **OCDE** publie une étude qui anticipe que 9% des emplois sont automatisables dans les pays de l'OCDE[1797] mais seulement 6% en Corée du Sud et 12% en Autriche. La France est dans la moyenne à 9%.[1798].
- **Juin 2016** : **Forrester** prévoit que 16% des emplois aux USA vont être automatisés d'ici 2025 que ce soit par de l'IA, des robots ou de l'automatisation classique[1799]. Cela sera compensé par la création de 9% d'emplois nouveaux (8,9 millions), générant une perte nette de 7% d'emplois. Ils prévoient que ce sont les emplois de bureau et administratifs qui seront les plus touchés. La création concernera la gestion des robots, les data-scientists et autres techniciens de la robotique et de l'IA ainsi que les curateurs de contenus. Mais ils n'anticipent pas la création d'emplois nouveaux indépendants des impacts technologiques. Ce sont les limites du modèle !

[1794] Voir Toil and Technology – Innovative technology is displacing workers to new jobs rather than replacing them entirely de James Bessen, 2015.

[1795] Voir From brawn to brains, the impact of technology on jobs in the UK, 2014 (16 pages).

[1796] L'étude est réalisée en collaboration avec Carl Benedikt Frey et Michael Osborne d'Oxford University. Petit détail : cette étude n'a évidemment pas anticipé le Brexit et son impact sur l'économie UK, qui si cela se trouve, sera plus grand que celui de l'automatisation ! C'est le syndrome du cygne noir !

[1797] Voir The Risk of Automation for Jobs in OECD Countries, mai 2016 (35 pages).

[1798] Voir OECD Principles on AI, mai 2019. L'OCDE s'appuie sur l'étude Frey/Osborne de 2013 qu'elle ajuste en corrigeant l'approche de ces derniers par une analyse de l'automatisation des tâches plutôt que des emplois, réduisant par conséquence les destructions d'emplois. L'OCDE n'indique pas d'horizon de temps ni n'évalue les créations d'emplois liées au développement d'innovations et de nouveaux services. En mai 2019, l'OCDE publiait une liste de 10 recommandations sur le déploiement de l'IA. Des définitions qui pourraient s'appliquer à tout le numérique voire à toutes les technologies : être bénéfique pour les gens et la planète, respecter les lois, les droits de l'Homme et la diversité, être transparente et auditable, etc.

[1799] Voir Robots, AI will replace 7% of US jobs by 2025, Forrester, juin 2016.

- **Septembre 2016** : **Forrester** récidive en prévoyant l'élimination nette de 6% des emplois aux USA[1800] d'ici 2021. Ce n'est pas très cohérent avec l'étude précédente publiée trois mois plus tôt, qui évaluait cette perte à 7% en 2025. Pourquoi une perte de seulement 1% sur quatre ans alors que c'est probablement la période pendant laquelle cette perte aurait tendance à s'accélérer ?

- **Mars 2017** : **MIT** et **l'Université de Boston** publient l'étude de Daron Acemoglu et Pascual Restrepo sous l'égide de l'organisme privé National Bureau of Economic Research sur l'impact de la robotisation sur le marché de l'emploi US[1801] entre 1990 et 2007. Ils déterminent que l'ajout d'un robot dans l'industrie détruit 6,2 postes du fait des répercussions sur l'ensemble de la chaîne de valeur des emplois. Mais cette étude est limitée à l'industrie manufacturière.

- **Avril 2017** : **Forrester** revoit ses prévisions de septembre 2016 à la baisse[1802]. Ils anticipent que d'ici 2027 l'automatisation va déplacer 17% des emplois aux USA et en créer 10%. On a donc toujours un solde de 7% mais à une échéance plus lointaine (2027 au lieu de 2025).

- **Mars 2017** : **PwC** publie une étude selon laquelle 38% des emplois US vont être automatisés d'ici 2030 dont 61% dans les métiers de la finance[1803]. On revient à des prévisions pessimistes. Les pourcentages équivalents sont de 30% sur UK, 35% en Allemagne et 21% au Japon. Ils tablent pour cela sur une généralisation des véhicules autonomes dès 2020. Ce qui créé évidemment un biais négatif énorme sur l'emploi, notamment pour les conducteurs de camions. Ils anticipent aussi une forte baisse de l'emploi dans la distribution. D'ailleurs, l'IA n'y joue pas un très grand rôle. Les douchettes de self-service dans les hypermarchés et le commerce en ligne est plus en cause.

- **Juin 2017** : **PwC** prévoit que l'IA va permettre d'augmenter le PIB mondial de 14% d'ici 2030, soient $15,7T (trillions de dollars)[1804]. Cette croissance serait due pour moitié aux gains de productivité et pour l'autre à l'évolution de la demande. Les gains seront plus forts en Chine (+26% de PIB) et de +14,5% en Amérique du Nord. L'Europe ne générerait que 9 à 12% de croissance du PIB et les pays émergents seulement 6% de croissance. Au-delà du fait que la méthode est probablement fantaisiste, on peut espérer que cette croissance ne sera pas indexée comme par le passé sur la consommation d'énergies fossiles.

  Cela rappelle les études du même genre et ordre de grandeur sur l'IOT qui étaient publiées entre 2013 et 2015. D'autres verront le jour sur la Blockchain et seront également redondantes.

- **Novembre 2017** : **Brookings Institution** analyse la numérisation de 545 métiers couvrant 90% des emplois américains[1805]. 70% des emplois créés depuis 2010 requièrent des compétences numériques modérées, et les emplois dont la dimension numérique est plus importante paient mieux.

- **Décembre 2017** : **Gartner** anticipe que l'IA va créer 2,3 millions d'emplois et en éliminer 1,8 millions aux USA d'ici 2020[1806] ce qui nous donne un solde positif de 500 000 emplois. Si la croissance économique US suit son cours actuel, il se pourrait que ces prévisions soient justes. Mais pas forcément que l'IA y soit pour quelque chose.

[1800] Voir Robots will eliminate 6% of all US jobs by 2021, septembre 2016 et The Top Emerging Technologies To Watch: 2017 To 2021, septembre 2016.

[1801] Voir Robots and Jobs: Evidence from US Labor Markets de Daron Acemoglu et Pascual Restrepo (91 pages) et Robots and Jobs Evidence from the US Labor Markets des mêmes auteurs, 2016 (69 slides).

[1802] Voir Forrester predicts automation will displace 24.7 million jobs and add 14.9 million jobs by 2027, avril 2017.

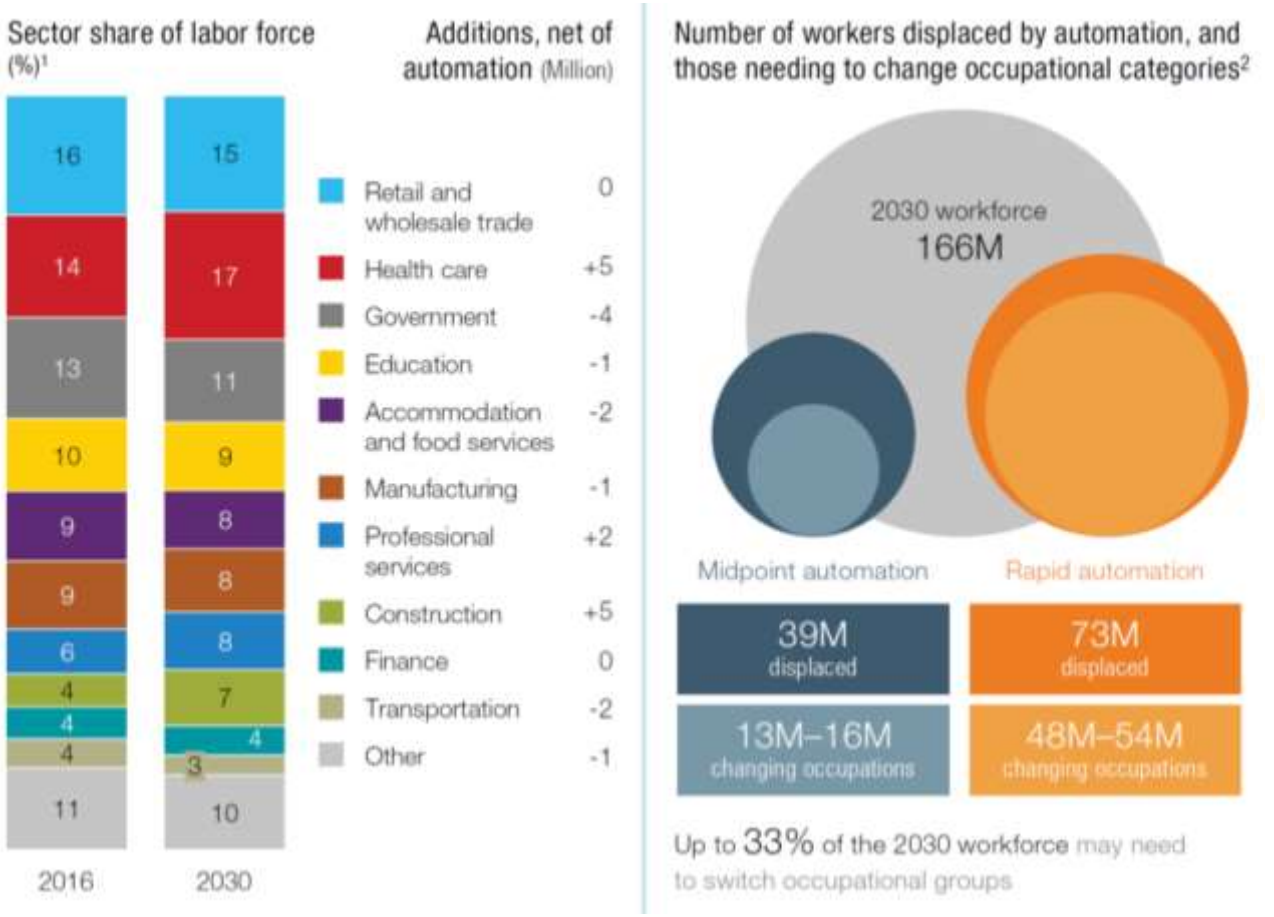
[1803] Voir Will robots steal our jobs? The potential impact of automation on the UK and other major economies, mars 2017 (19 pages).

[1804] Voir AI to drive GDP gains of $15.7 trillion with productivity, personalisation improvements, juin 2017.

[1805] Voir Tech Illiteracy Will Get You Fired Long Before Automation Does, novembre 2017.

[1806] Voir Gartner Says By 2020, Artificial Intelligence Will Create More Jobs Than It Eliminates, décembre 2017.

- **Décembre 2017** : **McKinsey** évalue la perte d'emplois à 7 millions aux USA en seulement trois ans. 14% des salariés devraient y changer de métier d'ici 2030 (schéma *ci-dessous*). C'est basé sur le fait qu'ils estiment que 15% des tâches seront automatisées d'ici à 2030, pour 9% en Inde et 24 % aux USA et en Allemagne et 29 % au Japon[1807]. McKinsey parle pudiquement de « displaced jobs » quand ils indiquent que des jobs vont être automatisés. Ils ne précisent pas vraiment ce que vont devenir les salariés qui occupent ces jobs qui vont disparaître car les jobs créés ne correspondent pas forcément à leurs compétences.



- **Janvier 2018** : **MIT** et l'**Université de Boston** mettent à jour du modèle mathématique de création/destruction d'emplois de Daron Acemoglu du MIT et Pascual Restrepo de la Boston University[1808].
- **Février 2018** : **PwC** publie une étude qui fait écho à celle de juin 2017, avec un modèle de prévision très sophistiqué qui couvre UK, Chine, Corée du Sud, Espagne et USA [1809]. Ils estiment la croissance du PIB mondial liée à l'IA de 13,8% d'ici 2030. L'IA va impacter 326 millions d'emplois à cet horizon (création + modifications).

[1807] Voir Jobs lost, jobs gained: Workforce transitions in a time of automation, novembre 2017 (160 pages). Voir aussi Jobs lost, jobs gained: Workforce transitions in a time of automation, septembre 2017 (14 slides) qui s'appuie sur l'exemple de l'arrivée des PC.

[1808] Voir Artificial Intelligence, Automation and Work de Daron Acemoglu du MIT et Pascual Restrepo de la Boston University, janvier 2018 (42 pages).

[1809] Voir The macroeconomic impact of artificial intelligence, PwC, février 2018 (78 pages).

- **Mars 2018** : **France Stratégie** publie un rapport sur l'impact de l'IA sur l'emploi en se focalisant sur trois marchés, les transports, la banque, la santé qui représentent 15% du PIB en France[1810]. C'est une étude qualitative qui résulte de l'audition d'environ 80 personnes. L'étude aboutit sur quelques recommandations portant sur la formation et sur le lancement d'un « chantier prospectif », le tout s'inscrivant dans le cadre de la mission Villani.

- **Juillet 2018** : **PwC** publie une nouvelle étude sur l'impact de l'IA et de la robotisation sur l'emploi au Royaume Uni[1811]. Selon elle, l'IA détruira à peu près autant d'emplois qu'elle en créera à un horizon assez lointain de 12 ans. Au Royaume-Uni, la création/destruction d'emplois portera sur 20% des emplois avec un solde positif de 200 000 emplois sur un total de 7 millions concernés, d'ici 2030, soient 0,5% des emplois. C'est très précis. Et probablement à côté de la plaque à cette échéance là. Ici, le Brexit est censé être pris en compte. L'étude décompose ce solde net d'emplois par région.

  Table 4.2: Estimated regional jobs impact of AI based only on variations in industry mix

Region	% of existing jobs (in 2017)			Number of jobs (000s)		
	Creation	Displacement	Net effect	Creation	Displacement	Net effect
London	22.0%	-19.7%	2.3%	1,297	1,159	138
South East	20.6%	-19.7%	0.8%	1,019	978	41
Wales	19.7%	-18.9%	0.7%	302	291	11
Scotland	20.2%	-19.6%	0.5%	558	544	15
South West	19.9%	-19.5%	0.4%	582	571	11
North East	20.0%	-19.8%	0.2%	239	237	2
East of England	20.4%	-20.3%	0.1%	648	646	2
North West	20.4%	-20.4%	0.0%	748	749	-1
West Midlands	20.1%	-20.4%	-0.3%	599	607	-8
Northern Ireland	19.4%	-19.8%	-0.4%	172	176	-4
Yorkshire and the Humber	20.0%	-20.4%	-0.4%	532	544	-12
East Midlands	19.5%	-20.7%	-1.1%	478	505	-27
Total	**20%**	**-20%**	**0%**	**7,176**	**-7,008**	**169**

  Source: PwC analysis

  On constate que ce solde est positif dans les grandes métropoles, surtout à Londres, et négatif dans les régions les plus pauvres comme le Yorkshire, les East Midlands et l'Irlande du Nord. L'IA va donc accélérer la concentration de la richesse dans les grandes villes.

- **Septembre 2018** : **World Economic Forum** (le WEF, organisation du forum de Davos) évoque la création de 47 millions d'emplois liés à l'IA d'ici 2022 représentant le solde de la création et la destruction de respectivement 122 et 75 millions d'emplois, le tout à l'échelle mondiale[1812]. Cela semble bien élevé même si 122 millions ne représentent qu'à peine 1,6% de la population mondiale.

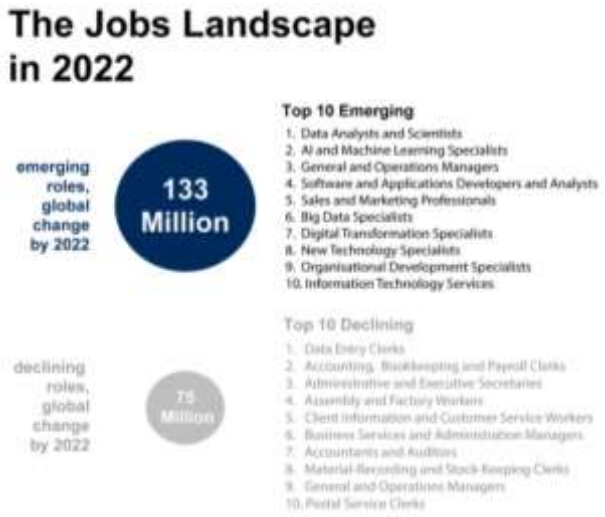
  Mais ces prophéties se réaliseraient qu'à une condition : que les gens soient bien formés. Ils indiquaient aussi que dans une autre étude, un seul métier a disparu en 60 ans : celui des liftiers d'ascenseurs[1813] ! Selon eux, un métier qui est partiellement et non totalement automatisé peut générer de la croissance dans l'emploi car il devient plus accessible à la clientèle. C'est le principe de la commoditisation.

[1810] Voir Rapport Intelligence Artificielle et Travail, France Stratégie, mars 2018 (90 pages).

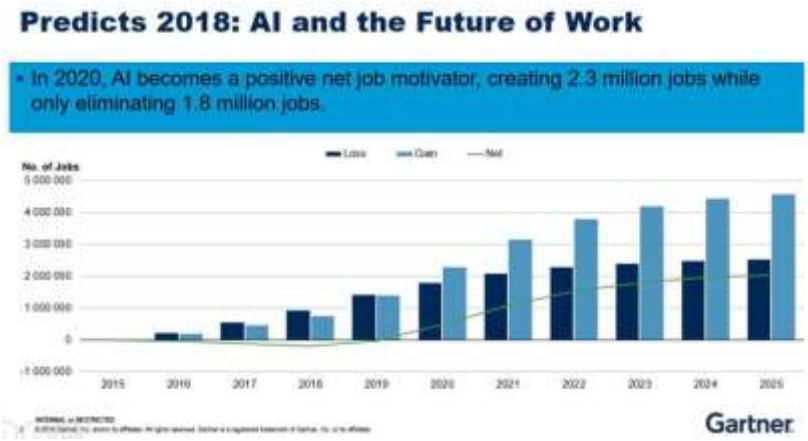
[1811] Voir UK Economic Outlook - Prospects for the housing market and the impact of AI on jobs, PwC, juillet 2018 (56 pages).

[1812] Voir 2022 Skills Outlook, World Economic Forum, septembre 2018.

[1813] Voir One. That's how many careers automation has eliminated in the last 60 years, mars 2017. Ils ont peut être oublié les dactylos qui n'ont peut-être pas entièrement disparues mais sont tout de même des plus rares dans les entreprises. Leur fonction est parfois cumulée avec celle de certaines assistantes ou assistants.



- **Octobre 2018** : le **Gartner Group** prévoit que le solde entre créations et destructions d'emplois lié au déploiement de l'IA commencera à être positif à partir de 2020[1814]. Sur quoi est-ce basé ? Sur des enquêtes clients. Ce qui ne veut évidemment rien dire surtout si les données ne sont pas ajustées comme dans les sondages politiques.



  Peu de dirigeants ont le courage d'affirmer dans ces enquêtes qu'ils sont prêts à dégraisser les effectifs grâce à l'automatisation et pour améliorer l'EBITDA de leur entreprise. Cela ferait désordre.

- **Novembre 2018** : une étude du FMI plaçait le risque d'impact de l'automatisation des métiers des femmes à 11% vs 9% pour les hommes[1815]. Cela concernait 26 millions de femmes dans 30 pays sur les 20 prochaines années. Les métiers occupés par des femmes auraient 70% de chances de plus d'être automatisés, impactant potentiellement jusqu'à 180 emplois occupés par des femmes.
- **Avril 2019** : nouvelle une étude de l'OCDE indiquait que la robotisation ferait disparaître 14% des emplois d'ici 2039. Le pourcentage augmente mais l'horizon également[1816]. Prévoir ce qui pourrait se passer économiquement vingt ans à l'avance est une belle gageure car cela ne tient pas compte des soubressauts géopolitiques, économiques, énergétiques, technologiques et maintenant... épidémiques qui pourraient intervenir.
- **Juin 2019** : un rapport de **McKinsey** sur l'automatisation de juin 2019 estimait que l'impact de l'IA sur l'emploi serait à peu près le même sur les femmes et les hommes, avec quelques varia-

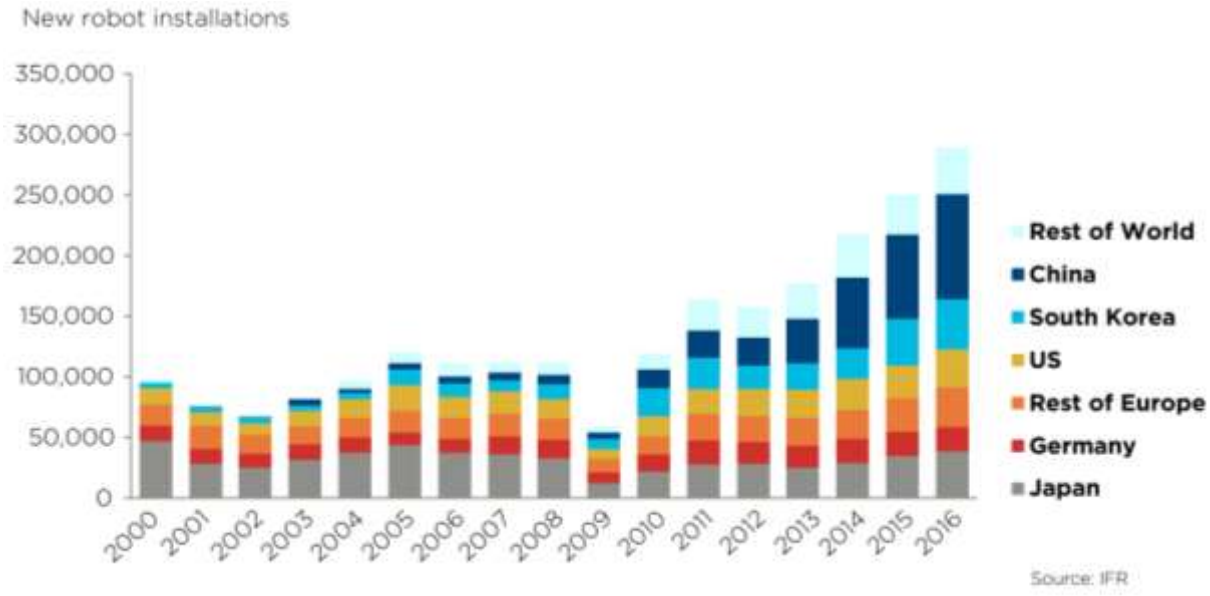
[1814] Voir Gartner : 2020, année charnière de l'intégration de l'IA en entreprise, octobre 2018.

[1815] Voir Women, Technology, and the Future of Work par Era Dabla-Norris et Kalpana Kochhar, novembre 2018.

[1816] Voir La robotisation devrait faire disparaître 14 % des emplois d'ici à vingt ans, selon l'OCDE par Marie Charrel, avril 2019.

tions d'un pays à l'autre[1817]. La répartition des métiers affectés aurait comme impact une plus grande difficulté d'adaptation des femmes par rapport aux hommes du fait des niveaux de qualification des métiers occupés par ces premières.

- **Juin 2019** : une étude de la société d'études privée **Oxford Exonomics** prévoyait que 20 millions d'emplois seraient remplacés dans le monde par des robots d'ici 2030, dont 1,5 millions aux USA, avec 14 millions de robots déployés en Chine dans ce laps de temps[1818]. Les régions industrielles les plus touchées seraient celles qui ont le plus de salariés faiblement qualifiés. Pour ce qui est de la France, l'étude considère que les régions à risque sont la Franche-Comté (du fait d'une forte activité industrielle), la Normandie, la Picardie, le Limousin et l'Auvergne, ceci dit, en utilisant le découpage des régions d'avant 2015, ce qui n'est pas très sérieux.



- **Septembre 2019** : c'était au tour d'**IBM** de publier une étude selon laquelle l'IA détruirait 120 millions d'emplois dans le monde en trois ans. Je m'étonne d'une telle précision sur un si petit volume à l'échelle mondiale et sur une période si courte[1819].

En février 2019, le fondateur de Y Combinator **Sam Altman** annonçait que le PIB mondial devrait augmenter de 50 % par an pendant quelques décennies. Il devrait arrêter le cannabis [1820]! Sur une décennie, cela générerait un PIB 57 fois plus gros qu'aujourd'hui. Coucou l'environnement ?

Au même moment, le directeur de l'éducation de l'OCDE estimait qu'il ne servait à rien d'apprendre aux enfants à coder car cela deviendrait une compétence obsolète. Encore un non scientifique qui n'a pas bien compris l'intérêt indirect de la programmation pour la réflexion, la résolution de problèmes et la conception de systèmes complexes[1821] !

[1817] Voir The future of women at work: Transitions in the age of automation par Anu Madg & Al, 2019 (168 pages).

[1818] Voir Robots could take over 20 million jobs by 2030, experts say, the trend could boost productivity and economic growth but also bring greater income inequality par Chloe Taylor, CNBC, juin 2019. Et la source de l'étude : How robots change the world – what automation really means for jobs and productivity, juin 2019 (64 pages).

[1819] Voir L'IA devrait détruire 120 millions d'emplois d'ici 3 ans par Véronique Arène, septembre 2019.

[1820] Voir A.I. may replace most of today's jobs, but this start-up investor sees global GDP increasing 50 percent a year within decades par Matt Rosoff, 2019.

[1821] Voir Le directeur de l'éducation de l'OCDE estime que c'est une perte de temps d'apprendre aux enfants à coder, car demain ce sera une compétence obsolète par Michael Guilloux, février 2019.

Analyses

Nombreuses sont les analyses qui cherchent à segmenter les métiers qui seront transformés de près ou de loin par l'IA, et parfois à vrai dire qui sont en ligne de mire prioritaire de transformation ou de remplacement par les technologies numériques en général.

De nombreuses précautions doivent être prises lors de prévisions de destruction/remplacement d'emplois résultant de l'automatisation :

- **Etudes et sondages**. Une bonne part des études sur l'emploi sont le résultat de sondages de dirigeants, pas d'analyse factuelle des technologies à venir et des rythmes prévisibles des innovations. Les enquêtes d'opinion sur le futur sont, comme en politique, à interpréter avec précaution.
- **Ne pas confondre métiers et tâches**. L'IA peut parfois automatiser certaines tâches mais pas les métiers dans leur ensemble. C'est le cas d'un cancérologue ou d'un ophtalmologue qui pourra bénéficier de l'apport de systèmes d'imagerie exploitant de la vision artificielle, mais qui auront toujours un rôle d'intégrateur des sources d'information sur le patient, des traitements et de la relation avec le patient dans la durée. L'automatisation amène à une recomposition complexe des métiers difficile à anticiper[1822].
- **Intégrer la dimension macro-économique**. Sont automatisables en priorité les métiers pratiqués de manière homogène sur des marchés larges, qui sont faciles à décrire et automatiser, où les ressources humaines sont soit rares soit trop chères, soit au mauvais endroit, avec des startups financées dans leur secteur et une réglementation favorable aux innovateurs, ce qui n'est pas le cas partout. Cela créé un filtre qui échappe à de nombreuses prévisions. Une analyse sur l'impact de la robotisation sur les emplois devrait porter sur leur structure. Les métiers sont très divers et fragmentés. Rien que dans la santé, on trouve des dizaines de types d'emplois et spécialités différentes. Il en va de même dans les services. Les startups s'attaquent en général en priorité à des cibles à la fois faciles et volumineuses, là où l'on peut générer une croissance exponentielle et de belles économies d'échelle au niveau mondial. Les kinésithérapeutes seront-ils remplacés par des robots bipèdes ? Probablement moins rapidement que les conducteurs de camions car ils sont moins nombreux, donc ne présentant pas les mêmes économies d'échelle potentielles ! Et l'automatisation du travail d'un kiné est plus complexe que celle d'un conducteur de camion. Par contre, nombre de métiers sont relativement protégés : ceux qui sont très manuels et difficiles à réaliser par des robots, les métiers créatifs et contenu relationnel[1823], ceux dont les tâches ne sont pas répétitives, ceux qui nécessitent des sens très pointus.

  Et puis bien sûr, ceux qui seront créés entre temps. Le monde des loisirs et du futile est assez prolixe de ce point de vue-là. Si l'on observe les nouveaux métiers créés depuis la fin de la seconde guerre mondiale, ils sont dominants dans ces catégories (tourisme, transports, médias, publicité, services divers, boutiques de tatouages, etc.).

[1822] Voir AI May Not Kill Your Job—Just Change It par Sara Harrison, octobre 2019, qui fait référence à The Future of Work: How New Technologies Are Transforming Tasks par Martin Fleming & Al, octobre 2019 (47 pages). L'étude porte sur l'évolution des tâches citées dans les offres d'emploi sur une période de plus de 10 ans. Elle indique que les tâches répétitives et automatisables sont progressivement remplacées par des tâches créatives et/ou relationnelles. Ils posent la question de l'évolution des emplois des classes moyennes qui correspondent aux plus grands enjeux de reconversion et de formation continue.

[1823] Apporter de l'amour en plus de l'IA ! Voir A blueprint for coexistence with artificial intelligence de Kai-Fu Lee, juillet 2017.

- **Pays émergents** : la majorité des prévisions ont aussi une fâcheuse tendance à se focaliser sur la situation aux USA et à ne pas adopter une approche mondiale du problème. Ils n'évoquent pas non plus le cas des fonctionnaires qui sont souvent les derniers à être robotisés car protégés par la lenteur de l'innovation dans les administrations et le manque de courage politique. L'impact de l'IA sur les pays émergents pourrait être encore plus sombre que pour les classes moyennes des pays développés. En effet, si la robotisation se poursuit dans l'industrie, elle supprimera des métiers d'exécution dans les pays émergents et transférera, dans une moindre mesure, de la valeur vers les pays développés, y compris ceux d'Asie (Corée, Japon, Taïwan et une partie de la Chine). L'automatisation des processus administratifs impactera de son côté les métiers de l'offshore, notamment en Inde.
- **Métiers en pénurie de compétences**. Dans le cas des médecins, l'automatisation ne réduira pas forcément l'emploi car le monde manque de médecins et notamment dans de nombreuses spécialités comme en ophtalmologie, en cardiologie ou en diabétologie. Les cancérologues ne sont pas non plus remplacés par IBM Watson. Ce dernier leur permet d'affiner leur diagnostic, leur prescription, et de les rendre plus personnalisés. Le métier de cancérologue est plus menacé par les progrès en médecine prédictive et en immunothérapies que par l'IA. A beaucoup plus long terme, les technologies permettant la prolongation de la vie en bonne santé pourraient cependant réduire le besoin en nombre de médecins, surtout si les maladies dites de longue durée sont éradiquées, cancers, diabètes et maladies neurodégénératives en premier. Certains actes de chirurgie seront aussi de plus en plus réalisés par des robots[1824]. Des phénomènes de vases communicants peuvent intervenir. Telle disparition entraîne la création d'emploi dans des secteurs connexes voire entièrement différents des métiers disparus. Pour ce qui est des radiologues, souvent présentés comme étant les plus menacés par l'IA dans les prévisions, ils sont 8736 en France sur un total de 325 870 médecins (données 2016). Si l'IA d'analyse des radios améliorait de 50% leur productivité, cela pourrait avoir un impact sur environ 4000 emplois. Donc environ 1,2% du total des métiers dans la santé. A ceci près qu'il faut se soucier du maillage territorial. Les vases communicants peuvent être à l'œuvre. On peut imaginer que le nombre de radiologues ne baisse pas mais que leur métier évolue pour intégrer une partie du travail des manipulateurs radios ! Les migrations de valeur peuvent être sérialisées ! Il faudrait aussi ajouter les 1500 anapats (anatomie pathologiques) en laboratoire dont une partie du travail peut-être assistée avec des techniques d'IA voisines de celles de la radiologie.
- **Création de métiers inconnus aujourd'hui**. Les prévisions s'accrochent trop souvent à notre vision actuelle des métiers, sans anticiper la création de métiers inconnus aujourd'hui, notamment dans le domaine des loisirs et de l'émotionnel avec les métiers de créatifs, les freelances et les métiers manuels dont les services à la personne[1825]. C'est aussi l'approche de l'économiste français Nicolas Bouzou, adepte de la prise de recul historique sur les craintes de destruction de l'emploi pour nous rassurer[1826]. Les prospectivistes ne sont d'ailleurs pas tous d'accord sur le sort qui sera réservé au métier d'enseignant et de docteur. Certains les voient entièrement remplacés par des robots et de l'IA, d'autres au contraire, non, car la relation avec les élèves et les patients devra rester humaine. C'est une question de perspective sur les aspirations humaines ! Or, si l'automatisation des métiers libère du temps et que le pouvoir d'achat des classes moyennes ne passe pas à la trappe (hypothèse…), alors, elle fera émerger de nouveaux besoins[1827].

[1824] Voir comme exemple : Surgical robot could sew you up better than a doctor de Jon Figas, avril 2016.

[1825] Voir Artificial Intelligence and Robotics and Their Impact on the Workplace produit par l'Iinternational Bar Association Global Employment Institute en avril 2017 (120 pages).

[1826] Notamment dans « Le travail est l'avenir de l'Homme », Nicolas Bouzou, 2017.

[1827] Voir AI Anxieties par Milton Ezrati, 2019. Qui n'a rien à voir avec moi !

Accenture PLC publiait en mars 2017 une étude terrain basée sur une enquête auprès de 1000 grandes entreprises qui évaluait la création d'emplois liés à l'automatisation d'autres emplois. Elle en liste quelques-uns, mais néglige les emplois créés dans d'autres domaines affectés par l'automatisation[1828].

The Jobs That Artificial Intelligence Will Create (Continued from page 15)

REPRESENTATIVE ROLES CREATED BY AI

Accenture's global study of more than 1,000 large companies identified the emergence of three new categories of uniquely human jobs.

TRAINERS	**Customer-language tone and meaning trainer**	Teaches AI systems to look beyond the literal meaning of a communication by, for example, detecting sarcasm.
	Smart-machine interaction modeler	Models machine behavior after employee behavior so that, for example, an AI system can learn from an accountant's actions how to automatically match payments to invoices.
	Worldview trainer	Trains AI systems to develop a global perspective so that various cultural perspectives are considered when determining, for example, whether an algorithm is "fair."
EXPLAINERS	**Context designer**	Designs smart decisions based on business context, process task, and individual, professional, and cultural factors.
	Transparency analyst	Classifies the different types of opacity (and corresponding effects on the business) of the AI algorithms used and maintains an inventory of that information.
	AI usefulness strategist	Determines whether to deploy AI (versus traditional rules engines and scripts) for specific applications.
SUSTAINERS	**Automation ethicist**	Evaluates the noneconomic impact of smart machines, both the upside and downside.
	Automation economist	Evaluates the cost of poor machine performance.
	Machine relations manager	"Promotes" algorithms that perform well to greater scale in the business and "demotes" algorithms with poor performance.

- **Automatisation sans IA** : il ne sera pas nécessaire d'atteindre un quelconque point de singularité où l'intelligence de la machine dépasserait l'homme pour que les tsunamis de l'emploi se produisent. Ils peuvent intervenir bien avant ! Et pour cause : bien des métiers d'exécution relèvent de tâches très répétitives qui sont sujettes à l'augmentation de l'automatisation dans un premier temps, sans passer par la case de l'AGI, l'intelligence artificielle générale, celle qui remplacerait totalement l'intelligence humaine, puis la dépasserait rapidement par la force démultiplicatrice des machines.

  Les prévisions oublient un autre phénomène induit par le numérique : le transfert du travail non pas seulement vers les machines mais aussi vers les clients, que l'on observe avec les distributeurs automatiques et caisses automatiques, le e-commerce, la SDA (sélection directe à l'arrivée) des centres d'appels, les chatbots (qui peuvent nous rendre aussi rapidement fous que les SDA) tout comme les banques et les assurances en ligne. Comme la valeur économique du temps des gens à faible revenu est faible, elle est absorbée en échange de services en théorie plus rapides[1829]. C'est un principe également courant dans l'économie collaborative, qu'elle concerne les professionnels (cas d'Uber, version VTC) ou les particuliers (Blablacar, Aibnb).

- **Démographie** : l'impact des progrès issus de l'IA sur la démographie sont-ils évalués ? Si la durée de vie s'allonge et le confort s'améliore, la démographie pourrait voir sa croissance ralentir, comme c'est le cas au Japon isolationniste depuis quelques décennies. Dans la réalité, elle restera inégale.

[1828] Voir The Jobs That Artificial Intelligence Will Create, mars 2017 (5 pages).

[1829] Le service n'est pas vraiment plus rapide mais on économise le temps de transport vers un point de service ou de vente.

Les technologies issues de l'IA ne se déploient pas à la même vitesse selon les continents et rien ne dit qu'elles éradiqueront les inégalités sur l'ensemble de la planète, surtout si le moteur de leur déploiement est hautement capitalistique[1830].

- **Timing** : on se trompe souvent sur le terme et même la nature des chamboulements. Ils sont généralement surestimés à court et moyen terme et sous-estimés à long terme, mais surtout mal appréhendés dans leur réalité technique et économique[1831]. Les études de cas mises en avant dans les ouvrages sur le futur de l'emploi collent souvent à l'actualité marketing du secteur de l'IA. Les livres parus après 2011 commencent presque tous par évoquer la victoire d'IBM Watson dans Jeopardy. A partir de 2013, ils passent aux prescriptions en cancérologie, l'une des applications commerciales de Watson, dont on découvrait les sérieuses limites en 2018. Depuis environ 2011, nous avons droit aux Google Car et autres avancées dans la conduite automatique. En 2016, ce sont les agents conversationnels (chatbots) qui sont devenus d'actualité, du fait de divers lancements comme dans Facebook Messenger.

- **Fact-checking** : en quelques années, les études de cas brandies en trophées peuvent perdre de leur substance. Il a été fait beaucoup de cas de la décision du Taïwanais **Foxconn** en 2011 de déployer un million de robots pour remplacer leurs travailleurs de ses usines en Chine qui demandaient des augmentations de salaire ou se suicidaient. Quatre ans plus tard, seulement 50 000 robots avaient été déployés[1832], ce qui ne présage rien de leur capacité à réaliser l'objectif annoncé mais illustre la difficulté à robotiser certains métiers manuels, même répétitifs. Dans cette abondante littérature sur le futur de l'emploi, les fondements scientifiques et technologiques des prévisions sont rarement analysés. S'y mêlent allègrement la science-fiction, la science et la fiction[1833].

- **Environnement** : les optimistes de l'innovation estiment que, grâce à l'IA, l'Homme sera capable de résoudre tous ces problèmes, presque d'un coup de baguette magique. En exagérant un peu, l'IA est devenue en quelque sorte la solution de sous-traitance ultime des sociétés procrastinatrices et des fainéants : ne nous attaquons pas aux problèmes qui fâchent et attendons que

---

[1830] Dans "The Demographics of Stagnation - Why People Matter for Economic Growth" de Ruchir Sharma dans Foreign Affairs, march-april 2016 selon qui la robotisation arrive à temps pour acompagner la baisse de la démographie dans les pays développés. Le Japon est un bon exemple : il cherche à produire des robots pour prendre en charge les personnes âgées car il n'y a pas assez de jeunes (ou d'immigrés) pour s'en occuper. Il y a moins de jeunes qui arrivent sur le marché du travail avec un effet retard de 18-22 ans sur cette baisse démographique. L'article ne le dit pas, mais la France a la particularité d'avoir une meilleure natalité qu'ailleurs en Europe. Mais ne la transforme pas pour autant en croissance et en emplois contrairement à de nombreux autres pays. Donc, la France est potentiellement plus vulnérable que d'autres pays à la robotisation des métiers.

[1831] Ainsi, dans Les robots veulent déjà nous piquer notre job d'Emmanuel Ghesquier (février 2016) qui commente une étude d'un certain Moshe Vardi de l'Université Rice du Texas, il est indiqué que l' "*on a pu voir avec les robots Pepper que certains robots pouvaient donner des conseils de gastronomie ou d'œnologie dans les supermarchés Carrefour ou qu'une boutique de téléphonie allait fonctionner à 100% avec des employés robotisés au Japon*". L'auteur qui relaie cela n'a pas du voir Pepper à l'œuvre car, au stade actuel de son développement, il est encore plus que brouillon ! J'avais même pu le constayer en 2014 dans une boutique Softbank dans le quartier Omotesando où ils commençaient à être déployés. Et ce n'est pas mieux dans toutes les démonstrations que l'on peut voir de ce robot dans différents salons professionnels. En y regardant de près, l'étude en question est un article publié dans The Conversation, Are robots taking our jobs. Il a bien du mal à faire le tri dans les évolutions de l'emploi aux USA entre ce qui provient de l'automatisation, de la globalisation et de la concurrence asiatique dans l'industrie manufacturière et même indienne, dans les emplois concernant les services informatiques. L'emploi a surtout migré géographiquement. Les emplois de perdus dans l'industrie aux USA et en Europe se sont retrouvés en Asie. C'est le "monde plat" de Thomas Friedman.

[1832] Voir Foxconn : les Foxbots peinent toujours à remplacer les humains, mai 2015.

[1833] Dans le top de l'exagération technique, nous avons par exemple **Tomorrowland** de Steven Kotler (2015), qui prédit monts et merveilles singularistes allant de l'intelligence artificielle générale (AGI) autoréplicable jusqu'au téléchargement des cerveaux dans un ordinateur : "*Yet it is worth noting that Moore's Law states that computers double in power* ***every twelve months*** *[…]. Biotechnology, meanwhile, the field where mind uploading most squarely sits, is currently progressing at* ***five times the speed of Moore's Law****. […] people alive today will live long enough to see their selves stored in silicon and thus, by extension, see themselves live forever.*". Nous avons donc une loi de Moore deux fois plus rapide dans les processeurs que dans la vraie vie (12 vs 24 mois) et des "biotechnologies" qui évoluent cinq fois plus rapidement que la loi de Moore, alors que cette vitesse ne concerne que le cas particulier de l'évolution du coût du séquençage de l'ADN, observée sur la période courte 2007-2011. Evolution qui s'est plutôt calmée les 5 années suivantes [1833]!

l'IA et la robotique fassent le boulot à notre place [1834]! C'en est presque un éloge du laisser-aller. Les ressources à notre disposition sont-elles infinies[1835] ? L'IA nous sauvera-t-elle assez rapidement d'un éventuel réchauffement planétaire accéléré dans les 30 ans qui viennent ? Les priorités de l'humanité pourraient en tout cas être sérieusement chamboulées.

- **Révoltes** : la *Ludditisation* des métiers n'est généralement pas évoquée dans les prévisions, du nom des Luddites qui résistèrent au début du 19ième siècle contre le développement des machines à tisser au Royaume-Uni. Tandis que la Reine Elisabeth I avait refusé l'octroi d'un brevet à William Lee en 1589, après son invention de la machine à tisser les bas, craignant de générer du chômage chez les ouvriers textiles, le gouvernement de sa Majesté avait décidé d'envoyer la troupe contre les ouvriers récalcitrants au progrès, entre 1806 et 1811. Un gouvernement élu par un parlement dominé par des entrepreneurs ! Quelles forces pourraient résister à l'automatisation des métiers ? Certains métiers ont-ils une meilleure capacité de résistance que d'autres, notamment par la voie de la règlementation ? Nous avons peu d'exemples résilients dans le temps !

Métiers

Segmentons donc les métiers qui seront impactés de près ou de loin par l'automatisation et par l'IA avec les métiers du passé qui disparaissent déjà, les métiers en train d'être automatisés qui vont disparaître ou être transformés, les métiers menacés, les métiers qui vont être créés et les métiers protégés[1836].

Métiers du passé

L'automatisation n'est pas nouvelle dans l'Histoire humaine. Elle a commencé il y a des millénaires avec l'agriculture, l'usage de bêtes de traits puis de tracteurs. Elle bat son plein depuis le 19e siècle et la première révolution industrielle. Elle s'est amplifiée avec l'avènement des outils numériques.

Les ouvriers de lignes d'assemblage ont déjà remplacés par des robots et le seront de plus en plus, surtout au gré du *in-shoring*, le rapatriement de la fabrication dans les pays occidentaux.

Les caissiers sont remplacés en partie par des automates de self-service et donc, par le travail gratuit des clients.

Les centres d'appels de taxis sont remplacés par des applications mobiles quand ce n'est pas la commande vocale. Aussi bien chez les sociétés de VTC à la Uber tout comme chez G7, le leader du marché traditionnel.

Métiers menacés

Les métiers en train d'être partiellement automatisés ne disparaîtront pas mais l'amélioration de leur productivité pourra réduire les effectifs[1837]. Cela concerne surtout les cols blancs et de nombreux

---

[1834] Voir AI: Helping, Not Threatening, Humanity par Hanno Schoklitsch, janvier 2020 qui évoque cette capacité de l'IA à aider à traiter le dérèglement climatique.

[1835] Deux ouvrages intéressants traitent assez bien de ces questions : **The beginning of infinity** de David Deutsch, qui défend un point de vue selon lequel l'infini et l'innovation sont intimement liés et qu'il ne faut pas de mettre des barrières à notre capacité d'innovation. Et puis **The infinite resource** de Ramez Naam qui fait un bilan circonstancié des défis qui se présentent pour gérer les ressources en apparence limitées de la planète côté énergie, agriculture et matières première. Il équilibre bien ces difficultés et les progrès techniques à venir qui permettront de les contourner.

[1836] A noter une différence entre emplois et métiers. Les emplois correspondent à un ensemble homogène de postes dans une structure et font appel aux mêmes compétences. Un métier est un ensemble d'emplois d'une même famille en termes d'activités et de compétences (source).

[1837] Voir L'impact de la révolution digitale sur l'emploi - Top 5 des métiers en voie de disparition de Erwann Tison, Institut Sapiens, 2018 (25 pages).

métiers de services, notamment dans les professions libérales administratives et dans la finance qui est de plus en plus automatisée[1838].

Dès lors qu'une tâche est répétitive ou qu'elle nécessite une prise de décision selon des règles assez simples, l'automatisation peut prendre le relai. C'est encore plus vrai si le marché est très concurrentiel et faiblement régulé.

L'IA automatise facilement les fonctions de reconnaissance d'image, d'où un impact sur un métier spécifique dans la santé : les radiologues. Cela ne veut pas dire qu'ils disparaîtront mais que leur productivité augmentant, ils pourront traiter davantage de patients. Cela concerne aussi les métiers de l'offshore comme les sous-traitants en Inde de processus d'entreprises qui pourraient être automatisés par les techniques de Robotic Process Automation que nous avions évoquées dans la rubrique sur ce vertical[1839]. On trouve dans ce créneau de nombreux métiers dans les banques et les assurances, dans la comptabilité, dans les fonctions de secrétariat. Côté travail manuel, la manutention pourrait progressivement passer à la moulinette de la robotisation dans les grands entrepots.

Métiers qui seront automatisés plus tard

Totalement ou partiellement comme dans la santé, le management, l'audit et même dans la recherche[1840]. L'activité de conducteur professionnel sera partiellement automatisée du fait de la généralisation des véhicules autonomes.

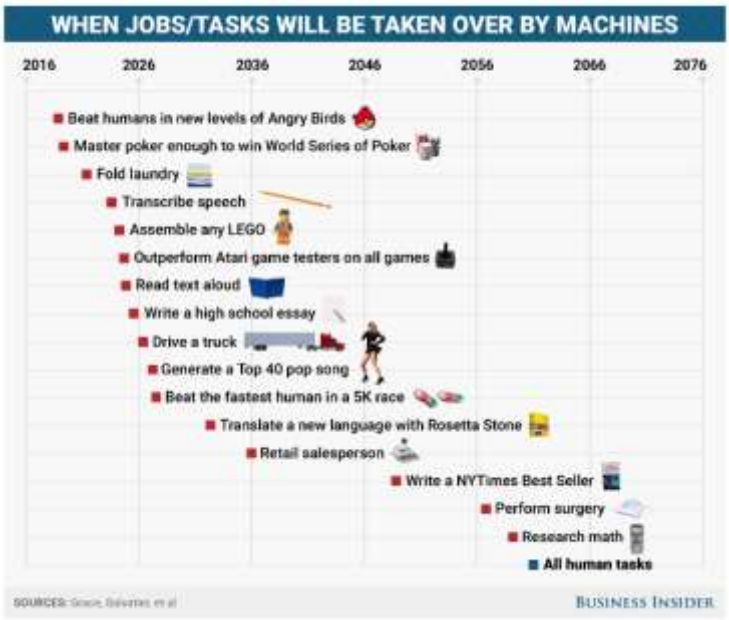
Mais on aura toujours besoin d'eux pour les chargements et déchargements et pour les trajets dans les zones denses. Il se pourrait bien que l'on ait besoin d'autant de conducteurs qu'aujourd'hui, comme pour les pilotes d'avions. Par contre, ils seront moins fatigués par leur travail du fait de l'automatisation de la conduite sur les longs trajets sur autoroutes.

[1838] Voir How my research in AI put my dad out of a job And what we are doing with the French government to prevent other people from losing theirs de Rand Hindi, fondateur de Snips.ai, mai 2017.

[1839] Voir Introduction to Robotic Process Automation, a primer, de l'IRPA, Institute for Robotic Process Automation, 2015 (35 pages).

[1840] Voir Les prochains paradigmes d'exploration scientifique seront peuplés d'Intelligences Artificielles, d'abord assistantes, elles deviendront collaboratrices, puis chercheuses d'Aymeric Poulain Maubant, octobre 2016.

Certains prospectivistes vont jusqu'à prévoir que le métier de développeur va lui-même être automatisé, ce à quoi je ne souscris pas du tout[1841]. Le schéma ci-dessous illustre les tâches qui vont être progressivement automatisées. Cela menacera donc les métiers qui les exécutent[1842]. Mais attention : tout ce qui dépasse la quinzaine d'année est totalement spéculatif. Il est très difficile de prédire les évolutions technologiques à de telles échéances. Il n'y a qu'à voir ce que l'on prédisait dans les années 1970 ou 1980 pour 2000, avec des taxis volants et autres fantasmes. Il faut se méfier des prévisions extrêmes[1843].



Métiers protégés

Différentes études inventorient les métiers qui ne sont pas menacés par l'automatisation, par l'IA ou la robotique[1844].

En voici quelques-uns :

- Les **métiers créatifs**, qui feront encore appel à des talents humains, même s'ils sont outillés par de l'IA pour étendre la palette d'expression. Dans tous les cas d'IA créative dont nous sommes abreuvés et qui font appel aux réseaux génératifs, le choix des outils et de la création reste l'affaire du créatif.
- Les **techniciens de maintenance**, même si leur activité est planifiée par des IA et de la maintenance prédictive et quand bien même on peut voir des démonstrations impressionnantes de robots de Schaft, maintenant chez Softbank Robotics. Cela concerne aussi les artisans dans le BTP, la plomberie et l'électricité.
- Les **métiers de services** qui touchent au corps : coiffeurs, manucures, kinésithérapeutes, etc. Ils ne sont pas prêts d'être remplacés par des robots. D'ailleurs, certains de ces métiers sont parfois plus menacés par la démocratisation d'outils exploitables par les consommateurs eux-mêmes.

[1841] Ce métier va évoluer comme il a évolué sur 50 ans, avec des transformations profondes, des différences accrues entre les développeurs de solutions utilisateurs assemblant des briques préexistantes et ceux qui les créent. Au même titre qu'aujourd'hui, la compétence pour faire évoluer le noyau de Linux n'est pas la même que pour créer un site en Wordpress avec des templates et des plugins. Il n'y a pas d'automatisation du métier, mais plutôt une stratification entre couches « hautes » et « basses » requérant des niveaux de compétences différentes.

[1842] Voir This is when robots will start beating humans at every task, juin 2017, qui fait référence à When Will AI Exceed Human Performance? Evidence from AI Experts, mai 2018 (21 pages).

[1843] Comme celles de Yuval Harari, cité dans Pour l'auteur de Sapiens, l'intelligence artificielle va détruire la plupart des emplois, Challenges, octobre 2017.

[1844] Comme These are the few jobs that robots won't take from us, de Michael Grothaus, août 2018.

- Les **professions de santé** sont en général faiblement menacées, surtout celles d'infirmiers/ières, les chirurgiens (même s'ils peuvent utiliser des robots), les médecins généralistes (qui vont pouvoir utiliser des outils de diagnostics plus nombreux), et de nombreux spécialistes. Si l'IA va certainement automatiser certaines tâches, elles portent sur des métiers qui sont en pénurie, comme les ophtalmologues. L'amélioration de leur productivité ne va donc pas les faire disparaître.
- Les **travailleurs sociaux** pour lesquels le contact humain est clé[1845].
- Les **enseignants** sont souvent mis en avant comme une profession protégée car essentielle. Et ce, malgré les nombreuses idées d'usage de l'IA dans l'éducation pour l'enseignant à distance et pour l'accompagnement des élèves. Nous avons vu pourquoi dans la rubrique de cet ebook sur l'éducation : la praticité de nombreuses solutions d'IA n'est pas évidente du fait de la grande fragmentation des activités de formation.

Métiers qui vont être créés.

Reste à étudier les métiers qui vont se développer voir être créés du fait de l'IA ou indépendamment de l'IA.

- Il y a bien entendu tous les **ingénieurs et développeurs en IA** et les métiers de support du numérique qui sont associés. Une bonne partie de l'écosystème numérique va basculer dans l'IA comme il a basculé dans l'Internet. L'IA sera rapidement une commodité qui perfusera dans tous les secteurs du numérique (startups, éditeurs de logiciels, constructeurs, sociétés de services, conseil, formation, entraîneurs de chatbots, labellisation de données).
- Les **métiers dans les loisirs** sachant que de nouveaux loisirs seront créés qui feront appel ou pas à l'IA. Dans ceux qui y feront appel, on peut évidemment penser aux applications de la réalité virtuelle et augmentée.
- Les métiers des **énergies renouvelables**. On le dit depuis longtemps et ce n'est pas lié à l'IA.

Revue de lectures

Voici quelques ouvrages de référence sur le sujet du futur du travail.

Rise of the robots and the threat of a jobless future (2016) de Martin Ford est un ouvrage bien documenté qui évoque un bon nombre des mécanismes macro-économiques des précédentes révolutions et crises industrielles, et de ce qui pourrait advenir dans le futur.

Sa thèse principale est que les révolutions numériques passées et à venir contribuent à réduire l'emploi dans les classes moyennes et à favoriser d'un côté l'émergence d'emplois de bas niveaux mal payés et de l'autre d'emplois de haut niveau bien payés. C'était déjà anticipé dans le rapport **Triple Revolution** produit en 1964[1846] pour l'administration de Lyndon B. Johnson. Ses auteurs s'alarmaient déjà sur risques les risques de l'automatisation, mettant en avant la difficulté de remplacer les emplois supprimés par la modernisation à un rythme suffisamment rapide. Il était très en avance sur son temps, alors que l'informatique n'en était encore qu'à ses balbutiements. Juste avant la sortie du mythique mainframe IBM 360, en 1965, c'est dire !

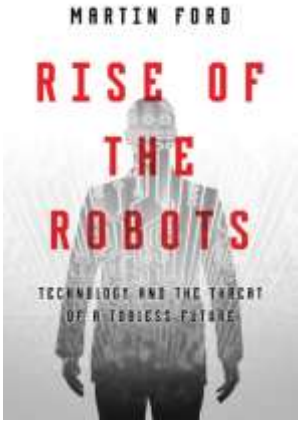
Aux USA, les 5% des foyers les plus aisées représentaient 27% de la consommation en 1992 et 38% en 2012. Les 80% les moins aisés sont passés de 47% à 39% dans le même temps.

Après la crise de 2008, le top 5% avait augmenté ses dépenses de 17% et le reste n'avait fait que rester au niveau de 2008. D'ou l'émergence de business comme Tesla qui cible, pour l'instant, surtout les 5% les plus riches.

[1845] Voir Yes, AI may take some jobs – but it could also mean more men doing care work, septembre 2018.

[1846] Voir The Triple Revolution, Libération, 1964.

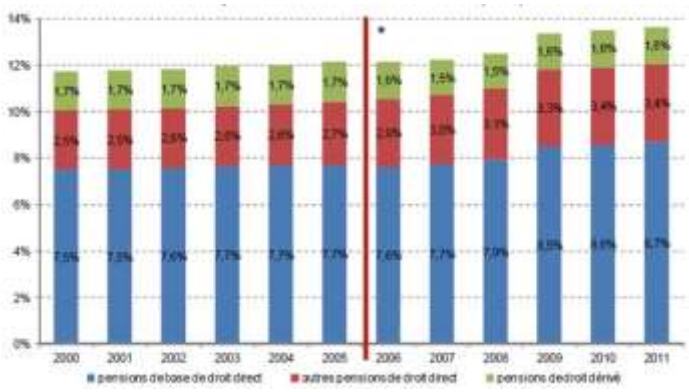
Les nouvelles entreprises issues du numérique sont automatisées dès le départ et ont moins de salariés. Elles profitent à plein de la productivité issue du numérique. Les exemples un peu éculés et trop généralisant de Whatsapp et Instagram sont mis en avant pour illustrer le point. On nous bassine un peu trop avec les $16B de "valeur" de Whatsapp générés par 55 employés, alors que lorsqu'elle a été acquise par Facebook, cette société n'avait quasiment pas de revenus.



Contrairement à l'après-guerre, les gains de productivité des deux dernières décennies sont allés non pas dans l'augmentation des salaires mais dans la baisse des prix, dans les salaires de métiers techniques qualifiés, et le capital s'orientant vers le financement des nouveaux investissements technologiques. Les technologies sont devenues un facteur d'inégalité au profit des technologues et des détenteurs de capital, tout du moins aux USA. La "finance" réalloue aussi les profits au bénéfice des plus riches. Plus un pays a un système financier développé, plus grandes seraient les inégalités.

Les profits des grandes entreprises ont augmenté sur 15 ans en proportion du PIB comme indiqué dans le schéma *ci-dessus* (source), qui correspond aux données US. Cette réallocation concerne 2,5% du PIB. Je me suis demandé où allaient ces profits. En pratique : un tiers alimente les fonds de pension, un autre tiers va dans les foyers, probablement avec des inégalités fortes de revenu, enfin, le reste va pour moitié chez des investisseurs internationaux, certains, aussi pour alimenter des fonds de retraite. Les méchants fonds spéculatifs (hedge funds) ne représentent que 4% de l'actionnariat des entreprises américaines [1847]!

Et si l'explication était donc toute simple : plus la population vieillit, plus les systèmes de retraites par capitalisation ont besoin de financement, donc de profits des grandes entreprises !



[1847] Voici la **source** du schéma correspondant.

En France, le régime général des retraites a vu son poids dans le PIB évoluer de 11,2% en 1990 à 13,8% en 2008, soient 2,6% de différence en pourcentage. Coïncidence ? Ne serait-ce pas finalement une solution différente au même problème ? A savoir, augmenter les charges sociales et taxes pour financer une retraite par répartition en lieu et place d'une augmentation du profit des grandes entreprises qui rémunèrent un système de retraite par capitalisation ? C'est probablement à moitié vrai et à moitié faux car les profits des grandes entreprises françaises ont aussi augmenté dans la même période. Mais comme les actions du CAC40 sont détenues par des investisseurs étrangers, il se trouve qu'ils alimentent aussi les systèmes de retraite de pays étrangers, notamment anglo-saxons qui en sont friands !

Le paradoxe est que la pénurie de compétences qualifiées ralentit ce phénomène de concentration de la valeur sur les plus riches ! Si on y pourvoyait plus rapidement, cela détruirait encore plus d'emplois faiblement rémunérés, et bien plus que d'emplois bien payés. La limitation des visas de travail pour les cadres qualifiés étrangers imposée par le congrès US créerait une inertie souhaitable pour protéger les emplois non qualifiés. En même temps, elle favorise l'offshore de métiers qualifiés en plus des métiers faiblement qualifiés qui sont déjà externalisés à l'étranger.

Autre point intéressant, l'auteur fait état des écueils des MOOCs, présentés comme la solution miracle pour l'enseignement. Deux études menées par l'Université de Pennsylvanie en 2013 et qu'il ne faudrait pas forcément généraliser montrent que les résultats d'étudiants ayant suivi des MOOCs étaient moins bons que ceux d'étudiants passant par des méthodes traditionnelles. Il ne faut certainement pas jeter le bébé du MOOC avec l'eau du bain de ces études. Les méthodes mixant MOOCs et enseignement IRL (in real life) sont probablement à favoriser.

L'ouvrage de Martin Ford met aussi en avant des opinions divergentes sur l'avenir de l'IA. L'expert en sciences cognitives **Gary Marcus** trouve que les performances récentes de l'IA sont survendues. Pour **Noam Chomsky**, qui s'est penché sur les sciences cognitives pendant 60 ans, on est encore à des millénaires de la création de machines intelligentes comme l'homme et que la singularité reste du domaine de la science fiction. Même opinion pour le psychologue cognitiviste **Steven Pinker**, le biologiste **P. Z. Myers** et même pour **Gordon Moore**. Il évoque aussi l'histoire de la National Nanotechnology Initiative lancée en 2000, qui survendait l'idée de créer des nano-machines au niveau des atomes et s'est ensuite rabattue sur des objectifs plus raisonnables.

Martin Ford évoque l'intérêt du revenu minimum qui est souvent présenté comme la solution pour traiter le problème de la disparition trop rapide d'emplois liés à la robotisation[1848]. C'est une sorte d'Etat providence générique poussé à l'extrême quand il n'est plus en mesure de créer les conditions d'une activité pour tous.

Ces débats font rage avant même que la richesse permettant de le financer ne soit créée et que de nouveaux métiers soient automatisés. La Finlande est parfois mise en avant comme validant le principe alors que le revenu minimum n'y était qu'à l'état expérimental entre 2017 et 2018 et n'avait d'ailleurs pas été reconduit ensuite, en avril 2018[1849].

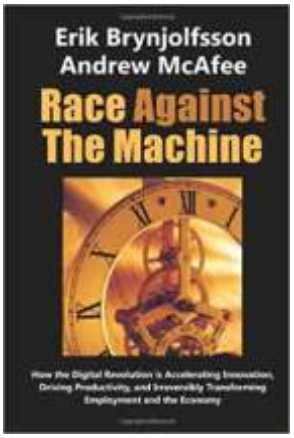
Les questions clés sont nombreuses. Quel est le niveau de ce revenu minimum ? Est-il là juste pour simplifier les systèmes existants de redistribution ? Comment est-il financé s'il est plus élevé ?

[1848] Ce revenu minimum (UBI : universal basic income en anglais) est tout de même proposé par Mark Zuckerberg, Richard Branson et Elon Musk. Voir Tax robots and Universal Basic Income de Ben Bloch, juillet 2018.

[1849] Voir Finland to end basic income trial after two years, avril 2018. Ce revenu minimum couvrait 2 000 personnes sans emploi âgées de 25 à 58 ans, avec un versement mensuel de 560€. La période de test n'était cependant pas suffisante pour en constater les effets, notamment en termes de réinsertion. Le gouvernement souhaitait revoir les conditions d'attribution et les associer à une formation professionnelle ou à l'acceptation d'un travail. Dans Pourquoi le revenu universel sera un interdit aussi fort que l'inceste de Laurent Alexandre, janvier 2018, insiste bien sur ce point : le RBI ne doit pas remplacer l'impérieuse nécessité de la formation pour adapter les sans emplois à l'économie de la connaissance. En France, le RSA est de 551€ pour une personne seule. On est donc au même niveau et on peut donc considérer que le revenu minimum est déjà en place en France. En fait, l'appréciation porte surtout sur le montant. Celui-ci est dans tous les cas en dessous du seuil de pauvreté. Et pour cause, celui-ci est souvent défini en proportion du revenu médian dans un pays. Il y a donc toujours une partie de la population qui est sous ce seuil.

Comment est-il différencié en fonction de la situation des foyers ? Comment évite-t-il de décourager les gens de travailler là où cela reste nécessaire ? Quel serait son impact si mis en place dans des pays et pas dans d'autres ? Quel impact sur les flux migratoires qui créent déjà une pression certaine ? Il existera toujours des inégalités marquées entre pays, en plus de celles qui existent entre milieux sociaux. Ce débat a démarré il y a plus de 11 millénaires avec les débuts de l'agriculture. Il s'est poursuivi avec toutes les autres révolutions technologiques et industrielles suivantes et n'est pas prêt de se terminer.

Dans **Race against the machine** (2012) et **The Second Machines Age**" (2014), Erik Brynjolfsson et Andrew Mcafee font les mêmes constats que le livre précédent sur la concentration de la richesse sur les 5% les plus aisés. Ils rappellent qui, si l'on considère aujourd'hui encore que les anciennes révolutions industrielles ont créé tant d'emplois, c'est parce que l'on a enlevé de l'équation les chevaux et autres bêtes de somme qui ont perdu leur utilité et ont disparu au passage, ou bien, ont été transformés en chair à steaks.

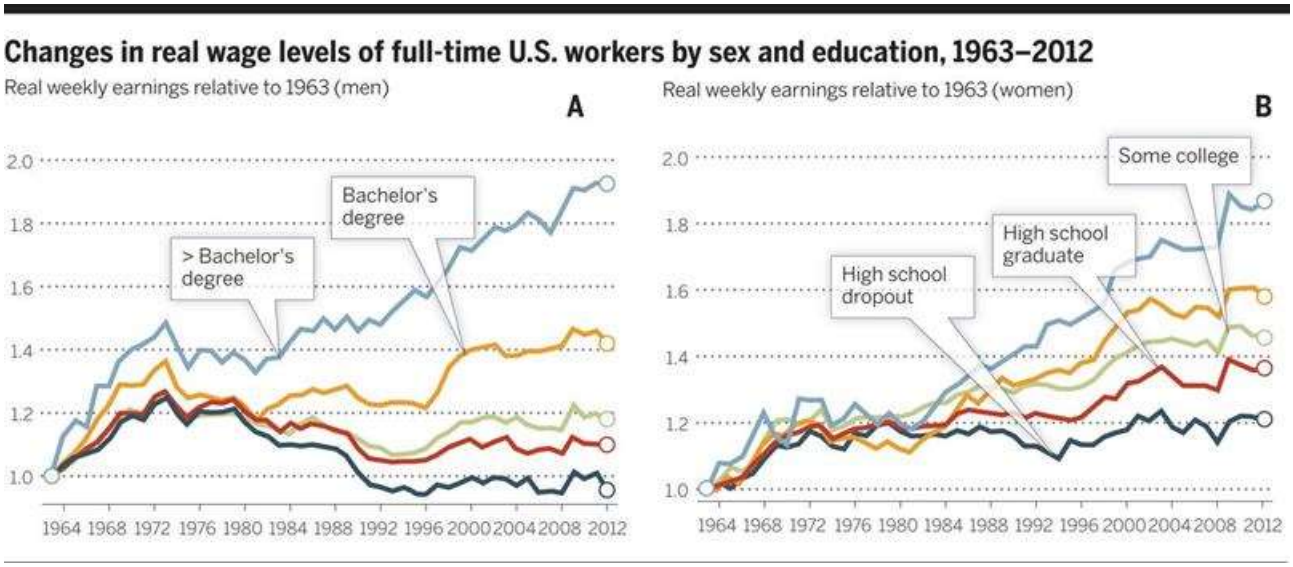


Ils étaient ce que sont aujourd'hui les travailleurs à bas salaire dont l'activité est en voie d'être automatisée, modulo les steaks. Le bilan écologique est aussi bien connu : c'est la terre qui a payé le prix de la croissance humaine !

Ils décrivent le scénario de l'offshore qui pourrait menacer l'emploi dans les pays à faible coût de main d'œuvre : les métiers délocalisés étaient les plus codifiables et donc, automatisables en priorité lorsque la technologie le permettra. Cela protège pour une part les pays occidentaux. A ceci près que les métiers codifiables non délocalisables pour des raisons physiques sont aussi automatisables. A contrario, le développement des robots réduit l'intérêt des délocalisations dans l'industrie. Il permet en théorie une relocalisation des usines, et la création d'emplois locaux de production, d'installation et de maintenance de robots ainsi que dans la chaîne d'approvisionnement.

Le scénario des auteurs met en avant les mêmes gagnants et perdants : les personnes à haut niveau de qualification contre les personnes faiblement qualifiées, les entreprises superstars à croissance exponentielle contre les autres, et enfin le capital contre le travail. Il s'appuie sur le fait que, ces dernières décennies, les salaires ont déjà augmenté pour les personnes les plus qualifiées et baissé pour les moins qualifiées (source des graphes *ci-dessous*).

On pourrait ajouter à cette analyse la possibilité d'un ajustement de la population mondiale en fonction des glissements de valeur provoqués par la robotisation. Quelle serait l'influence de la robotisation sur la natalité ? Et surtout de la prolongation de la durée de la vie, sans même parler de vie éternelle. Plus la longévité augmente, comme au Japon, plus la natalité est en baisse. A court et moyen terme, cela résout le problème de l'emploi par le vide. Mais une société vieillissante peut enclencher son déclin inexorable. L'impact d'un éventuel revenu de base ne serait pas neutre. Avec lui, la démographie n'irait plus naturellement à la baisse.



Les deux auteurs, qui sont de la MIT Sloan School of Management, proposent un plan d'action en quatre points qui s'inspire en partie des propositions du **rapport Triple Revolution** de 1964 :

- Investir dans l'**éducation**, en payant mieux les enseignants, en les rendant responsables, et attirer aux USA les immigrants qualifiés. Côté cursus, ils recommandent d'investir dans la créativité, dans l'identification de tendances et dans la communication complexe. Ils font remarquer que l'homme associé à la machine sont plus puissants qu'une machine seule[1850]. Donc, associer la créativité et la maîtrise de l'usage des technologies reste une belle protection. Ils considèrent que tous les métiers qui requièrent à la fois de la créativité et une forte sensibilité motrice ne sont pas prêts d'être automatisés (cuisiniers, jardiniers, réparateurs, dentistes). Les auteurs font aussi preuve de bon sens en rappelant que notre imagination est limitée pour prédire les emplois du futur. On n'anticipe pas assez la nature des problèmes existants et à venir qui vont générer leurs propres métiers.
- Développer l'**entrepreneuriat** : l'enseigner comme une compétence dans l'ensemble de l'enseignement et pas seulement dans les meilleures business schools, réduire les réglementations qui ralentissent la création d'entreprise, et créer un visa pour les entrepreneurs. Ils recommandaient aussi d'encourager les innovations d'organisation et du travail collaboratif pour exploiter ce qu'il reste d'utilisable du temps et des compétences des gens inoccupés.
- Développer l'**investissement** dans l'innovation, la recherche et les infrastructures, notamment dans les télécommunications. Un grand classique des pays modernes comme des pays émergents.
- Côté **lois et fiscalité**, ne pas alourdir la législation du travail. Rendre les embauches plus attractives que la robotisation des métiers au niveau des charges sociales et taxes, ce qui rappelle une bonne partie de la politique de l'emploi en France, qui ne nous réussit pas si bien. Ne pas réguler les nouvelles activités. Réduire les subventions aux emprunts immobiliers et les réallouer à l'éducation et à la recherche. La propriété immobilière a tendance à réduire la mobilité géographique. Réduire les subventions directes et indirectes aux services financiers. Réformer le système des brevets et réduire la durée d'application du copyright.

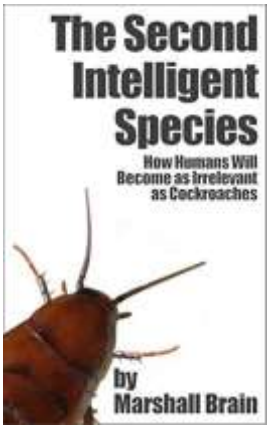
[1850] Dans **Human+Machine**, Paul Daugherty et H. James Wilson insistent sur le couplage homme plus machine et non pas le remplacement de l'homme par la machine. L'ouvrage est bien commenté par Yves Caseaux dans Réinventer les processus et les applications avec l'Intelligence Artificielle, Yves Caseaux, septembre 2018.

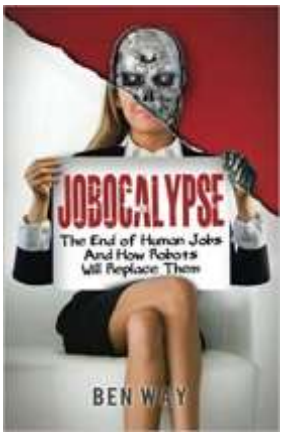
Enfin, ils ne recommandent pas de créer une allocation universelle mais plutôt un crédit d'impôt pour les bas revenus (negative income tax) dans la lignée d'une proposition de Thomas Paine qui date de 1797 au Royaume-Uni. Pourquoi valoriser le travail ? Parce que, quelle que soit sa nature, en plus de pourvoir à nos besoins, le travail traite deux nuisances : l'ennui et le vice (Voltaire), sans compter les couches hautes de la pyramide des motivations de Maslow.

C'en est presque un plan "à la Macron" : favorisons l'entrepreneuriat et tous les problèmes sociétaux se régleront d'eux-mêmes. Un peu trop classique !

The Second Intelligent Species: How Humans Will Become as Irrelevant as Cockroaches (2015), de Marshall Brain, grossit le trait en annonçant que les scientifiques sont en train de créer une seconde espèce intelligente, les robots et l'IA, qui va nous dépasser et supprimer la majorité des emplois. Les premiers touchés seront les millions de camionneurs, les vendeurs dans la distribution de détail, dans les fast foods et le BTP. C'est un darwinisme technologique provoqué par l'Homme, qui se fait dépasser par ses propres créations.

Le reste est de la non-science-fiction, tablant sur une intelligence artificielle qui régulerait les comportements humains néfastes, comme ceux qui affectent l'environnement. Les emplois non qualifiés disparaîtraient à la fin des années 2030, ce qui semble un peu rapide au vu de la progression de la robotique. Au passage, l'auteur fournit une explication du fameux paradoxe de Fermi selon lequel il est bizarre qu'aucune civilisation extraterrestre ne nous ait approchés à ce jour. "Officiellement", diraient les conspirationnistes. L'IA développée par ces civilisations serait comme la nôtre : une fois qu'elle serait satisfaite par ses réalisations et par l'équilibre ainsi généré, elle n'aurait pas besoin d'explorer le reste de l'univers.

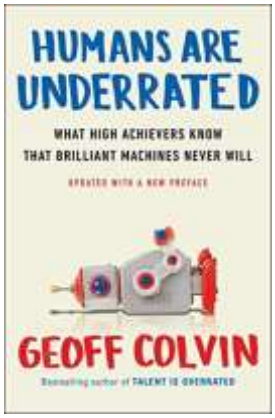




Jobocalypse (2013) de Ben Way, que je n'ai pas lu (désolé…), part du principe que nous sommes *déjà* envahis par les robots et que la disparition d'emplois liée à l'automatisation est une histoire ancienne. Il anticipe que même les métiers les plus qualifiés seront remplacés par des robots car ils s'autoalimenteront. Les scénarios envisagés vont de révolutions provoquées par les sans-emplois à des initiatives gouvernementales de formation massive les concernant. On dira que l'on préférera le second scénario au premier même si c'est un peu court !

Quand au Rapport **Global Catastrophic Risks 2016** de la Global Challenges Foundation, il intègre l'IA dans les risques systémiques que l'humanité et la planète pourraient rencontrer, au même niveau que les conséquences du réchauffement climatique et les pandémies naturelles ou artificielles. Les risques évoqués ne concernent cependant pas les conséquences sur l'emploi mais plutôt la perte de contrôle de l'IA par l'Homme. Cette évaluation est mise à jour dans l'édition 2017 du rapport[1851].

[1851] Voir Global Catastrophic Risk 2016 (55 pages) et Global Catastrophic Risk 2017 (67 pages).



Enfin, **Humans Are Underrated: What high achievers know that brilliant machines never will** (2015) de Geoff Colvin met en avant de son côté l'opportunité de remettre au goût du jour les qualités humaines dans les métiers : l'empathie, l'intuition, la créativité, l'humour, la sensibilité et les relations sociales. Une manière de différentier clairement les machines et l'homme.

C'est aussi l'approche proposée par Bruno Teboul dans **Robotariat – critique de la robotisation de la société**, paru au printemps 2017, qui associe philosophie, économie, prospective et humanisme pour envisager un monde équilibré où l'IA et les data sciences ne sont pas mises au pilori, et sont utilisées pour faire avancer la société, et où la place de l'homme et de la nature est préservée. Il y propose un revenu universel de reconversion à vie financé par les entreprises qui automatisent, de taxer le trading à haute fréquence et de développer une vie plus écologique[1852].

Elle était reprise par Dov Seidman dans **Harvard Business Review** en 2014[1853] pour qui les métiers du futur sont les métiers qui ont du cœur ! C'est une belle conclusion, même si frisant quelque peu l'utopie[1854].

Parades

Maintenant que le problème est posé, comment ne pas être remplacé par de robots et de l'intelligence artificielle ? Après l'uberisation qui intermédie les métiers de service, la robotisation peut-elle automatiser ces mêmes métiers ? La robotisation serait-elle la forme ultime d'ubérisation ?

Quelques pistes sont bien connues et déjà citées dans les livres évoqués ci-dessus : choisir des métiers où la créativité, l'initiative, les émotions, l'empathie et l'humanité sont importantes et adopter les nouvelles technologies qui rendent plus productif. Et ne tombons pas dans le panneau des annonces tonitruantes d'IA créatives ! Ce sont des outils pour les créatifs, pas des IA qui se passent de créatifs.

Comme avec toute nouvelle technologie, de nouvelles formes de créativité humaine verront le jour. Les outils de l'IA permettent aux créatifs de tout poil de se poser de nouvelles questions. Un scientifique peut ou pourra explorer la connaissance et l'état de l'art plus facilement. Un chercheur pourra faire des hypothèses et les vérifier plus facilement. Un ingénieur pourra simuler encore plus aisément ses créations. Un urbaniste pourra évaluer l'impact d'un projet.

[1852] Cet ouvrage très instructif permet de découvrir ou de redécouvrir de nombreux auteurs clés de ces différents domaines. Il décrit avec recul des phénomènes récents comme l'uberisation. Et il partage quelques défauts avec ce document : des parties de deux à trois pages qui survolent de nombreuses thématiques, un panorama peu-être un peu trop large et pas assez profond, et un jargon pas forcément accessible comme ces « Prolégomènes à une herméneutique des NBIC » (en langage courant, on dirait peut-être « Prélude à une interprétation des NBIC ». !

[1853] Voir From the Knowledge Economy to the Human Economy de Dov Seidman, novembre 2014.

[1854] Voir aussi cette alerte sur la tendance des outils numériques à réduire les véritables interactions humaines Eliminating the human et à limiter notre prise de risques : L'impossible voyage connecté, ou comment le numérique a étouffé le sentiment d'aventure.

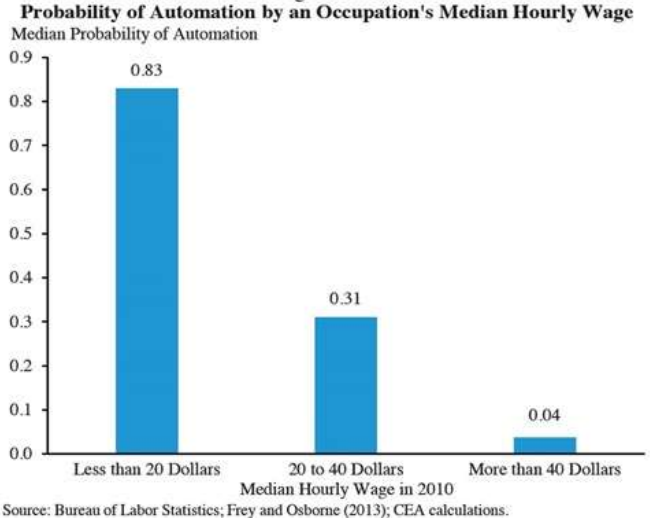
Un marketeur pourra faire de même avec des hypothèses produit et marché. L'IA permettra de créer de nouveaux outils de compréhension de l'existant et de simulation de nouveaux projets dans tous les domaines.

L'abondance des données exploitable par les IA ne fait pas tout ! Il faut savoir se poser les bonnes questions pour les exploiter !

C'est une constante dans l'innovation : dans presque tous les métiers, l'automatisation et la robotisation ne sont jamais totales. Elle nécessite une supervision humaine. Il faut donc s'approprier les outils de cette supervision, voire les créer soi-même ! Donc, de préférence, maîtriser à la fois des métiers traditionnels et les technologies numériques qui peuvent les transformer.

Malheureusement, les sciences et technologies n'attirent pas tant que cela les jeunes et notamment en France, comme une enquête mondiale réalisée par **Randstad** le montre[1855].

Figure 5-15
Probability of Automation by an Occupation's Median Hourly Wage



Source: Bureau of Labor Statistics; Frey and Osborne (2013); CEA calculations.

A contrario, il faudra de préférence éviter les métiers répétitifs, routiniers ou à faible degré de créativité et d'initiative et simples d'un point de vue moteur. Ce sont ceux qui présenteraient le plus grand risque d'automatisation.

Le schéma *ci-dessus* issu du **Rapport Economique du Président US 2016**[1856] rappelle que les métiers à bas salaire, donc en général à faible qualification, sont les plus menacés par l'automatisation.

Dans **The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution** publié en janvier 2016 par le World Economic Forum[1857], les auteurs prévoient que les deux tiers des enfants en école primaire d'aujourd'hui exerceront un métier qui n'existe pas encore. Ils y vont un peu fort car l'échéance n'est pas si lointaine.

Ils prévoient que 7,1 millions d'emplois administratifs disparaitront d'ici 2020, et que seulement 2 millions d'emplois seront créés dans les technologies (aux USA). Par contre, des emplois devraient être créés pour combler une partie du trou dans l'énergie, les nano-biologies et le divertissement, et ceux des commerciaux subsisteront.

Et oui, les emplois de l'avenir seraient surtout ceux dont le contenu émotionnel sera le plus dense, comme expliqué dans **Les 10 compétences clés du monde de demain**[1858].

D'un point de vue stratégique, on peut intuitivement privilégier l'enseignement supérieur, la recherche et l'entrepreneuriat dans les domaines scientifiques et technologiques qui génèrent ces automatisations.

Il vaut mieux créer ou adopter les outils de l'automatisation que de n'en subir que les effets, comme décrit dans **How To Avoid Being Replaced By A Robot** paru dans Fast Company en avril 2016[1859].

[1855] Voir Un Français sur quatre conscient d'être remplacé par un robot, avril 2016.

[1856] Voir Economic Report of The President 2016, février 2016.

[1857] Voir The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution, janvier 2016.

[1858] Voir Les 10 compétences clés du monde de demain, mars 2016.

[1859] Voir How to avoid being replaced by a robot ?, avril 2016.

On pourra aussi favoriser les enseignements pas trop spécialisés et assez diversifiés. Et enfin, ne pas oublier d'exceller dans ce qui fait de nous des Hommes, de belles machines biologiques douées d'émotions.

Politique par l'IA

Les métiers de la sphère marchande ne sont pas les seuls à être impactés par l'IA. Celui de politique l'est tout autant, même s'il n'a pas vocation à être une activité à temps complet dans toutes les démocraties.

Le cerveau fonctionne très souvent par analogies et la connaissance de l'Histoire influe sur les décisions des politiques, sauf lorsqu'ils ne connaissent pas du tout l'Histoire comme Donald Trump. L'IA n'utilise pas encore massivement le raisonnement par analogies. Il répond surtout en fouillant dans de vastes dépôts de connaissances et pour croiser quelques informations structurées. Mais sait-on…

Politique fiction

Est-ce qu'une IA pourrait indiquer : si tu envahis tel pays dans telle et telle circonstance, voici ce qui a le plus de chances de se produire en suivant les leçons de l'histoire connue. Et voici ce qui permettrait d'éviter le pire ! Une IA pourrait-elle guider un exécutif dans les choix de politiques à la fois rationnelles et irrationnelles ?

On apprend souvent du passé pour (mieux ?) décider du futur. Comment des décisions politiques complexes influencent la sphère économique qui agit à la fois de manière rationnelle et irrationnelle aux événements ?

Autre difficulté à surmonter pour l'IA, mais pas insurmontable : comment tenir compte d'un adversaire qui agit de manière non rationnelle ? La plupart des algorithmes d'IA sont conçus de manière rationnelle. Exemple : comment réagir quand l'une des parties agit de manière irrationnelle, tel un Saddam Hussein en 1990/1991, voire lorsque les deux parties sont irrationnelles avec ce même Saddam Hussein et Georges W. Bush en 2003 ?

Je m'étais aussi demandé en 2013[1860], pour les 50 ans de l'assassinat de JFK, si un système de type Watson ne pourrait pas un jour analyser toute la littérature sur le sujet et pondre une synthèse voire résoudre l'énigme qui est bien plus complexe qu'une simple théorie du complot style 9/11 ou sur les chemtrails. L'analyse des faits et mystères de l'histoire pourrait probablement gagner de ce genre de système. Mais l'intérêt économique de la chose est plutôt marginal !

Est-ce que les organisations politiques et des Etats peuvent se faire eux-mêmes disrupter par de l'IA ? Evitons l'expression "uberiser" qui est à la fois trop précise et trop vague.

Il y a bien l'initiative **Watson for President** mais elle est un peu légère car construite comme une opération de communication d'IBM[1861]. Elle visait un peu à la manière de Coluche en 1980, de faire élire Watson comme nouveau président américain en 2016. En indiquant que cela permettrait à la Maison Blanche de prendre des décisions rationnelles. C'est confondre un peu rapidement l'outil de la prise de décision (POTUS) et l'outil d'aide à la prise de décision (Watson et/ou le staff du Président et son administration).

Un président fait déjà appel à de nombreux experts pour prendre ses décisions, en particulier dans la diplomatie, les négociations internationales et le pilotage du bras armé des USA. Il a aussi besoin de pas mal d'aide et de tacticiens pour faire voter des lois par le congrès qui est souvent récalcitrant, même lorsqu'il est du même bord que lui.

[1860] Voir Les technologies et l'assassinat de JFK, novembre 2013.

[1861] Voir http://watson2016.com/.

On l'a vu pour l'Affordable Care Act (Obamacare) lors du premier mandat de Barack Obama. Mais avec Donald Trump, on peut songer à l'avantage qu'il y aurait eu à élire une IA au lieu d'un tel personnage !

La première question à se poser sur l'usage de l'IA concerne les élections dans les démocraties. Les dernières grandes élections, notamment américaines, ont montré la force à la fois des réseaux sociaux et de la propagation d'idées véhiculant du rêve (Obama, Sanders) ou des peurs et angoisses, et les fameuses fake news (Trump). L'élection de Donald Trump a montré comment la manipulation des opinions pouvait faire basculer de peu une élection[1862].

Que ferait l'IA pour améliorer un tel processus ? Elle collecterait des volumes gigantesques d'informations ouvertes sur ce qui se dit et s'écrit, sur ce que font les électeurs, sur leurs réactions à des discours antérieurs, sur les analyses biométriques (de la captation de pouls avec une montre, des mouvements oculaires avec des capteurs de Tobii, de l'EEG pour la mesure de l'activité cérébrale, etc), sur l'économie ou sur les médias.

Elle les analyserait alors au point de permettre la création de programme politiques appliquant soit la **démagogie ultime** (celle qui fait gagner les élections mais qui est inapplicable ou qui, si appliquée, mène à une catastrophe[1863]) soit la **démagogie utile** (celle qui fait à la fois gagner les élections et aller dans un chemin non catastrophique et responsable). Le tout en étant conforme à une idéologie de base d'un parti politique donné, avec son système de valeur (partage, social, économie, croissance, environnement, fiscalité, justice, école, immigration, selon les cas). Voilà de beaux défis d'optimisation sous contraintes[1864] !

Political Speech Generation

Valentin Kassarnig
College of Information and Computer Sciences
University of Massachusetts Amherst
vkassarnig@umass.edu

Abstract

In this report we present a system that can generate political speeches for a desired political party. Furthermore, the system allows to specify whether a speech should hold a supportive or opposing opinion. The system relies on a combination of several state-of-the-art NLP methods which are discussed in this report. These include n-grams, Justeson & Katz POS tag filter, recurrent neural networks, and latent Dirichlet allocation. Sequences of words are generated based on probabilities obtained from two underlying models: A language model takes care of the grammatical correctness while a topic model aims for textual consistency. Both models were trained on the Convote dataset which contains transcripts from US congressional floor debates. Furthermore, we present a manual and an automated approach to evaluate the quality of generated speeches. In an experimental evaluation generated speeches have shown very high quality in terms of grammatical correctness and sentence transitions.

Des tentatives de ce genre ont déjà été vaguement lancées. Valentin Kassarnig, chercheur à l'Université Amherst du Massachusetts, a présenté début 2016 un premier **générateur de discours politique** basé sur de l'IA[1865], et qui dépasse les générateurs de pipeau déjà bien connus.

Le résultat reste assez rustique et focalisé sur le langage, pas sur la construction d'un programme politique qui se tienne. La solution est même diffusée en open source[1866] ! Malheureusement, en politique plus qu'ailleurs, l'adage selon lequel le contraire de l'IA est la bêtise naturelle s'applique parfaitement. Cette dernière est même plutôt efficace électoralement !

[1862] Hillary Clinton a devancé Donald Trump en vote populaire de 2,8 millions de voix mais a perdu le college des grands électeurs pour 78 000 électeurs dans quatre *swing states*, qui avaient fait l'objet d'un ciblage particulier de fake news dans les réseaux sociaux. J'avais fait une analyse chiffrée de cette élection dans L'origine et les conséquences de l'élection de Donald Trump, novembre 2016.

[1863] Voir The future of democracy in the AI era par Sukhayl Niyazov, novembre 2019 qui évoque l'impact de l'IA : la création de bulles d'informations, la tyrannie de la minorité liée à la démagogie portant sur les « swing voters », les fameux indécis, le côté émotionnel et faiblement rationnel des élections, la vulnérabilité de ces émotions aux manipulations. Mais bon, pas besoin d'IA pour manipuler ces émotions. Les politiques, surtout extrêmes, savent déjà très bien le faire !

[1864] Voir A.I. Will Mark A Turning Point in the History of Politics par Sukhayl Niyazov, février 2020, qui véhicule un tel optimisme, un peu béat, sur ce genre de scénario.

[1865] Voir Political Speech Generation de Valentin Kassarnig, 2016 (15 pages).

[1866] Voir https://github.com/valentin012/conspeech.

Après les élections se pose la question de la gestion. Est-ce que l'IA permettrait de préparer des choix censés mis ensuite dans les mains d'électeurs dans le cadre de démocraties plus participatives ? Est-ce que l'IA permettrait de bâtir des politiques économiques dignes de ce nom ? Est-ce l'IA permet d'intégrer les complexes relations sociales dans la société ? D'anticiper les réactions des citoyens aux nouvelles lois et réglementations, notamment fiscales ? Est-ce qu'elle permettra de gérer les conflits ? Est-ce qu'elle pourrait permettre d'accélérer la justice ? D'éviter les erreurs judiciaires ? De réformer les systèmes de santé au fil de l'eau des progrès technologiques ? Je n'en sais rien. Il n'y a pas beaucoup de chercheurs qui planchent sur ces questions !

Certains indiquent toutefois qu'une IA impliquée dans le processus apporterait un peu de rationalité et serait capable de prendre des décisions non basées sur le côté obscur des émotions[1867]. Un sondage de 2019 indiquait que les citoyens seraient enclins à plus faire confiance à une IA qu'à des politiques pour les diriger. Mais c'est évidemment faire fi de l'intervention humaine pour créer cette IA et des biais des données qui l'alimenterait[1868].

Mais l'IA peut servir à peaufiner des décisions techniques opérationnelles. C'est ce que permettent les solutions de simulation de la ville intelligente, permettant d'optimiser l'urbanisme des villes en fonction des contraintes de mobilité avec un mix mouvant de moyens de transports individuels et collectifs et la prise en compte des questions prastiques (perte de temps), énergétiques (consommation d'énergie fossile) et environnementales (bruit, pollution).

C'est l'objet du projet collaboratif Smartgov qui sert à concevoir et simuler des politiques urbaines. Il utilise probablement des techniques à base de réseaux multi-agents. Le projet associe le LIRIS-CNRS et Naver Labs Europe à Grenoble.

Les systèmes d'aide à la décision politique pourraient-ils faire appel à de l'IA intensive ? Y compris lorsqu'il s'agit d'évaluer la position et l'attitude des autres parties prenantes, des agents économiques ou des chefs d'Etat ? Est-ce qu'une IA permettrait à un POTUS[1869] de gérer de manière optimale la relation conflictuelle avec Vladimir Poutine, les bras de fer avec les Chinois, ou de résoudre pacifiquement les divers conflits du Moyen-Orient ? Ou à un Président français de se dépatouiller de la situation en France ?

On a bien vu des films de Science Fiction mettant en scène des personnages liés à l'IA comme dans "Her" et "Ex Machina", mais pas encore dans la politique fiction. Ca ne saurait tarder vue l'imagination débridée des scénaristes ! Un « House of Cards » avec un « Special Assistant to the President » qui soit une IA à commande vocale ferait son effet et nous éloignerait de la présidence-réalité actuelle !

On en est encore loin. Ce qui démontre par l'absurde que l'AGI (Artificial General Intelligence) n'est pas pour tout de suite. Mais gare à vos fesses les politiques ! La démocratie participative pourrait prendre un visage inattendu !

Est-ce qu'une IA pourrait améliorer la prévision sur les résultats des élections par rapport aux sondages ? Le cas le plus complexe est celui des élections présidentielles américaines avec leur système alambiqué de grands électeurs et de swing states qui sont par définition imprévisibles.

Un chercheur de New York avait tenté de prédire le résultat de la présidentielle en se contentant d'analyser les flux sur Twitter qui n'a joué de rôle dans ce genre d'élection que depuis celle de 2012 et encore. Au final, il concluait juste pour ses prévisions concernant la présidentielle de 2020 que Joe Biden devait obtenir l'avantage en vote populaire mais qu'il était difficile de se prononcer du côté des grands électeur[1870].

[1867] Voir Should Politicians be Replaced by Artificial Intelligence? Interview with Mark Waser, 2015.

[1868] Voir Encore plus effrayant que l'IA Militaire, l'IA Président de la République ? par Jean-Paul Oury, 2019.

[1869] POTUS = President Of The United States.

[1870] Voir 2020 election: Artificial Intelligence has chosen a winner - but there's a catch, octobre 2020.

Une autre IA avait prédit la défaite de Trump en 2020, sans connaître son opposant démocrate et avant la crise du Covid, ce qui revenait à aller un peu vite en besogne[1871]. En 2016, une IA d'origine indienne avait prédit l'élection de Donald Trump en analysant les flux dans les réseaux sociaux mais pour combien d'autres IA ayant prédit celle d'Hillary Clinton [1872]? Il faut prendre toutes ces prévisions à base d'IA avec des pincettes, surtout lorsqu'elles sont annoncées après les événements, avec un évident biais du survivant. Elles n'analysent que la superficie des choses et des opinions et calculent des corrélations entre divers paramètres, mais pas au point d'être bien plus précises que les habituels sondages, qui eux-mêmes sont biaisés autant par les réponses des électeurs votant aux extrêmes et ne l'avouant pas toujours que par les indécis qui, par nature, peuvent facilement changer d'opinion et au dernier moment[1873].

Politique réalité

L'IA est en fait déjà utilisée en politique dans quelques contextes sporadiques pour ou contre les démocraties.

Le machine learning a été utilisé pour la segmentation et le ciblage d'audiences clés dans diverses élections par l'anglais **Cambridge Analytica**, surtout en 2016[1874].

La société a notamment analysé les profils d'électeurs sur Facebook dans les *swing states* lors de la présidentielle américaine de novembre 2016. Cela a permis ensuite à d'autres équipes périphériques à la campagne de Trump de cibler des populations de swing states avec des fakes news fabriquées par des sites conspirationnistes « alt-right », sans compter l'effet de relais des bots créés par des équipes financées directement ou indirectement par la Russie poutinienne et mafieuse.

Si on allait plus loin, on pourrait imaginer l'utilisation de réseaux de neurones génératifs fabriquant de fausses vidéos de personnalités, pour influencer l'opinion qu'en ont les électeurs. Pas besoin d'AGI pour y arriver !

Il suffit d'exploiter les technologies existantes et d'avoir de mauvaises intentions. Le mal, c'est l'Homme ! Yuval Harari s'inquiète à juste titre de l'usage de l'IA par des dictatures[1875].

- L'exploitation sémantique des analyses terrains réalisées par les volontaires de « En Marche », par **Proxem** (2007, France). Cette société avait produit juste après l'annonce de la candidature d'Emmanuel Macron une longue présentation montrant quels mots clés ressortaient des enquêtes terrain. Cela avait l'air d'exploiter des techniques assez basiques de traitement du langage.

[1871] Voir Donald Trump will lose the 2020 election, concludes unique prediction model that nailed 2018 midterm results, par Douglas Perry, juillet 2019.

[1872] Voir Une intelligence artificielle avait prédit la victoire de Trump (et de ses prédécesseurs) par Alexis Orsini, novembre 2016.

[1873] Voir le très bien documenté IA et élections : quelques succès, beaucoup d'échecs par Benoit Deshayes, février 2020.

[1874] L'origine, l'histoire et les méthodes de Cambridge Analytica sont bien documentées dans la présentation Uses and abuses of AI in election campaigns de Alistair Knott (85 slides). La société a été créée en 2013 et financée par l'investisseur Bob Mercer, un ultra-conservateur, cofondateur du site d'information Breitbart. Cambridge Analytica avait comme VP un certain Steve Bannon, passé pendant 7 mois à la Maison Blanche comme conseiller spécial de Donald Trump. Pour la présidentielle 2016, la société exploitait diverses sources de données acquises, notamment auprès d'Acxiom et Experian. Cela leur a permis de constituer une base nominative de plus de 200 millions d'américains adultes avec 5000 données associées. Ils ont même fait des expériences d'A/B Testing sur des messages TV s'appuyant sur leurs données de profiling. Voir aussi Cambridge Analytica, the shady data firm that might be a key Trump-Russia link, explained de Sean Illing, octobre 2017. J'écrivais ceci en octobre 2017. Depuis, le scandale de Cambridge Analytica a éclaté au printemps 2018 aux USA du fait de la capture de données privées sur Facebook de dizaines de millions d'utilisateurs par Cambridge Analytica en 2015 et 2016. Cela a entraîné une commission enquête du Sénat US puis la fausse fermeture de Cambridge Analytica devenu Emerdata. Voir The power players behind Cambridge Analytica have set up a mysterious new data company, mars 2018 et Cambridge Analytica dismantled for good? Nope: It just changed its name to Emerdata, mai 2018. Voir aussil'excellent documentaire d'Arte : Comment Trump a manipulé l'Amérique, octobre 2018. On a appris ensuite que Cambridge Analytica avait repéré les préférences politiques des électeurs via des corrélations assez poussées utilisant leurs vêtements. Voir Fashion's role in Cambridge Analytica's 'cyber warfare,' according to Christopher Wylie, 2018.

[1875] Dans Why Technology Favors Tyranny de Yuval Harari, octobre 2018.

- La tentative de prévision de ce que le Congrès US pourrait voter, par un professeur de l'université de Vanderbilt, J.B. Ruhl et le développeur John Nay dans PredictGov[1876]. Le site n'est plus en ligne et n'a probablement pas prévu le vote de John McCain au Sénat américain contre la suppression d'Obamacare fin juillet et fin septembre 2017. Il ne pouvait pas non plus prévoir l'issue du vote de confirmation du juge Brett Kananaugh à la Cour Suprême par ce même Sénat en octobre 2018.
- L'**Alan Turing Institute** anglais ambitionne de modéliser les sources de conflit pour les prédire sur une période de cinq ans[1877]. Il a même fait un appel du pied pour lancer une initiative dans ce domaine auprès de l'ONU.

Tout cela reste encore artisanal. Mais ce n'est peut-être que la face visible d'un gros iceberg méconnu. Mais qui régulera les usages de l'IA par les politiques ?

Cela devient moins artisanal lorsque des gouvernements de pays non démocratiques font appel à des acteurs technologiques occidentaux pour mettre en place des moyens de surveillance électronique et de filtrage assistés par de l'IA[1878].

Des affaires récentes de ce genre impliquant Google et Microsoft ont fait réagir des ONG ainsi que des salariés de ces entreprises. Ils réagissaient de la même manière lors de la signature de contrats entre leur entreprise et le Pentagone aux USA.

Les réseaux génératifs pourront aussi servir à créer des articles de propagande conspirationnistes qui aggrègent des bases existantes de propagandes.

Et si l'IA pouvait argumenter ? L'**IBM Project Debater** est une IA qui intervient dans des débats avec des humains sur des sujets variés. C'est un projet dans la lignée d'IBM Watson gagnant au Jeopardy en 2011. Pour l'instant Project Debater n'est pas encore à la hauteur des meilleurs humains mais il fait des progrès constants. L'enjeu était de faire changer une audience d'avis en 15 minutes d'argumentation sur une question de société (le subventionnement des écoles maternelles). Il mettait en évidence qu'une argumentation rationnelle est généralement insuffisante pour emporter un auditoire. La conviction compte tout autant. Dans ce cas, les débatteurs humains avaient même réussi à faire pencher la balance vers la position contraire à celle de Project Debater. Il ne suffit pas de piocher de manière probabiliste dans un stock de 10 millions de phrases pour convaincre [1879]!

Politiques de l'IA

L'agitation des peurs autour de l'IA en a fait un véritable sujet politique. Le phénomène n'est pas nouveau en soi mais l'est par son ampleur. D'habitude, les pouvoirs publics s'emparent de sujets technologiques avec un retard de phase chronique. Ici, ils sont quasiment en avance de phase, en tout cas, relativement aux menaces de l'IA.

La posture politique suit une gradation relativement classique entre tenants d'une innovation Schumpétérienne libérale qu'il ne faut pas tenter de ralentir et ceux d'un État régulateur et protecteur cadrant les usages, l'éthique des affaires et l'économie en général. La première est dominante aux USA tandis que la seconde l'est en Europe et surtout en France où l'étatisme ne faiblit jamais, et résiste fort bien aux alternances politiques.

L'IA est aussi une révolution industrielle et les États ont compris qu'il ne fallait pas louper le coche. Des deux côtés de l'Atlantique, les États sont des plus prolixes en rapports et plans autour de l'IA.

[1876] Voir Artificial intelligence tries to make sense out of the mess that is Congress, avril 2017.

[1877] Voir Retool AI to forecast and limit wars, Alan Wilson & Al., octobre 2018 (3 pages).

[1878] Voir How artificial intelligence systems could threaten democracy par Steven Feldstein, août 2019.

[1879] Voir Ce robot d'IBM spécialisé dans les discours reste moins convaincant qu'un humain, février 2019 (vidéo).

Barack Obama avait été interviewé par Joi Ito dans Wired en août 2016 et articulait déjà une vision claire des enjeux autour de l'IA[1880]. Juste après, son administration a avait produit deux rapports en fin de mandat, le premier The Administration's Report on the Future of Artificial Intelligence, publié en octobre 2016 après une consultation publique faisait quelques recommandations élémentaires : l'IA devrait servir à améliorer le bien public, les gouvernements devraient l'utiliser, l'IA devrait compléter et non pas remplacer les hommes, et l'usage des véhicules autonomes devrait être régulé. S'en suivit Artificial Intelligence, Automation, and the Economy en décembre 2016, qui anticipait des bouleversements sur le marché de l'emploi qui peuvent être absorbés pour peu que les efforts adéquats soient lancés côté formation et qu'une réflexion sur la répartition de la valeur ait lieu.

Un comité juridique du Parlement Européen publiait en février 2017, un court rapport European civil law rules in robotics, plaidant pour la création d'un cadre juridique encadrant l'usage des robots mais en s'opposant à une taxe sur les robots.

La France a suivi de peu avec la publication par le gouvernement d'un rapport et de l'initiative France IA en mars 2017, produits en deux mois concomitamment avec un rapport des assemblées portant principalement sur les questions sociétales et d'éducation soulevées par l'IA.

Peu de temps après l'arrivée d'Emmanuel Macron à l'Élysée, le gouvernement a confié au Député Cédric Villani la mission qui porte son nom pour « *étudier les actions nécessaires pour permettre à la France et à l'Europe d'être à la pointe de l'économie fondée sur l'IA* » et pour identifier les moyens de moderniser l'État avec l'IA. Le rapport était remis fin mars 2018.

Au Royaume-Uni, le Parlement a lancé un appel à consultation sur ces mêmes sujets en juillet 2017 alors que sa Commission Science et Technologie avait déjà publié un premier rapport en octobre 2016, assez succinct avec 44 pages, Robotics and artificial intelligence.

D'autres rapports ont été produits par divers groupes de pression sur les États, comme en Australie, où le cabinet de conseil AlphaBeta publiait en août 2017 The automation advantage (44 pages, *ci-contre*) qui dénonce le retard des entreprises australiennes dans l'adoption de l'IA et fait miroiter un potentiel économique de $2,2T (trillions = 1000 milliards), probablement un peu surestimé au regard du PIB actuel du pays qui est de $1,2T même si ce potentiel économique est étalé sur quinze ans.

L'une des inquiétudes partagées par les politiques et certains milliardaires de la tech est le déséquilibre que l'IA pourrait créer dans la fabrique sociale des sociétés, détruisant à petit feu la classe moyenne qui assure sa stabilité[1881]. Se posent plein de questions autour de l'IA : formation, développement des startups, organisation de la recherche, adoption de l'IA par les entreprises, régulation et protection de la vie privée.

Rapports et plans

Nous allons creuser ici l'Histoire des rapports sur l'IA en France, étalée entre 2017 et 2018 avec le premier plan France IA (mars 2017), le rapport des assemblées (mars 2017), le rapport de la Mission Villani (mars 2018), le rapport de France Stratégie sur l'IA et le travail (mars 2018), le plan de l'Académie des Technologies (avril 2018) et le plan de la région Ile de France (octobre 2018).

[1880] Voir Barack Obama, neural nets, self driving cars, and the future of the world, dans Wired, octobre 2016. Sa position était bien documentée et articulée. On se met à fantasmer sur une interview comprenant exactement les mêmes questions, mais posées à Donald Trump. L'A/B testing de POTUS serait saisissant !

[1881] Voir How Technology wrecks the middle class de David Autor et David Dorn, aout 2013 et Why governments need to respond to the Fourth Industrial Revolution de Iain Klugman, septembre 2018.

plan France IA du gouvernement
mars 2017

rapport de l'OPECST
Sénat+Assemblée
mars 2017

rapport Mission Villani
Assemblée Nationale
mars 2018

France Stratégie
IA et travail
mars 2018

Région Ile de France
Plan IA 2021
octobre 2018

Plan France IA

Je faisais un tour d'horizon de ce plan et rapport de 350 pages dans un article publié en mars 2017[1882] et dont voici un résumé légèrement actualisé.

Ce plan décrivait une recherche fondamentale assez dispersée et centrée sur la recherche publique et d'un manque de transferts technologiques, ce qui n'est pas une spécificité de l'IA. Il évoquait le rôle stratégique des données qui alimentent l'IA. On y trouvait une cartographie de la recherche française en IA complété par celle des entreprises l'utilisant.

Il comprenait des propositions d'orientation de cette recherche, notamment dans l'IA symbolique sous toutes ses formes qui complète l'IA connexionniste qui domine l'univers du deep learning pour la vision et le traitement du langage. Le rapport détaillait aussi les stratégies de grands acteurs américains et chinois ainsi que les plans lancés par différents gouvernements dans le monde.

Le plan proposait de se focaliser sur sept marchés : l'automobile, la relation client, la finance, la santé, les énergies renouvelables, la robotique et l'éducation numérique[1883].

Le plan d'action prévoyait notamment la candidature à un *projet phare de technologie émergente* de l'Union Européenne (« FET flagship ») du type du Human Brain Project, sur l'IA, pouvant être financé à hauteur de 1Md€, le lancement d'un programme IA dans le cadre du Plan pour les Investissements d'Avenir (PIA 3), le financement d'une infrastructure mutualisée de calcul de puissance en IA pour la recherche et la création d'un consortium public-privé sur l'intelligence artificielle (devenu le Hub France IA en septembre 2017).

Le plan oubliait, semble-t-il volontairement, les composants électroniques, qui sont un domaine où les opportunités autour de l'IA sont significatives avec au moins deux technologies clés : les processeurs neuromorphiques et les processeurs quantiques généralistes qui pourraient transformer radicalement le paysage informatique autour de l'IA dans les deux décennies à venir.

[1882] Que j'ai commenté ici au moment de sa publication : Les hauts et les bas du plan France Intelligence Artificielle, mars 2017.

[1883] Un an plus tard, le rapport Intelligence artificielle : État de l'art et perspectives pour la France, février 2019 (333 pages), d'Atawao Consulting réalisé pour le compte de la DGE, du Commissariat général à l'égalité des territoires et de l'association TECH'IN France, revenait un peu plus en détails sur certains de ces marchés avec un ensemble de recommandations. Il est dommage qu'il contienne des erreurs grossières historiques et techniques (page 11 : Kasparov n'a pas perdu en 1992 mais en 1997 contre Deep Blue, page 27 : Another Brain ne conçoit pas un processeur quantique, page 51 : confusion sur la définition du deep learning, …).

Rapport des assemblées

Publié quasiment simultanément au plan France IA en mars 2017, le rapport de l'Office Parlementaire d'Évaluation des Choix Scientifiques et Technologiques[1884] « **Pour une intelligence artificielle maîtrisée, utile et démystifiée** » complétait le plan France IA de l'exécutif en abordant surtout la dimension sociétale et réglementaire.

Nourri par des déplacements à l'étranger (USA, Royaume-Uni, Suisse) et de nombreuses rencontres, le rapport démarre avec un très bon panorama de l'histoire de l'intelligence artificielle qui remet pas mal de pendules à l'heure, même s'il contient quelques perles. On y découvre aussi les 12 laboratoires impliqués dans l'IA au CNRS (GREYC, IRIT, LAAS, LAMSADE, LATTICE, LIF, LIG, LIMSI, LIPN, LIRMM, LORIA, LRI, TIMC) qui totalisaient alors environ 300 chercheurs permanents. Le rapport comprenait surtout une quinzaine de propositions regroupées en trois parties reprenant les éléments du titre du rapport :

Côté **maîtrise**, on sentait poindre des relents de principe de précaution, avec des oxymores tels que la volonté d'éviter une régulation de la recherche en intelligence artificielle tout en voulant favoriser une IA sûre, transparente et juste via des chartes, de la formation à l'éthique de l'IA, la création d'assurances spécifiques aux robots et la création d'un Institut National de l'Ethique de l'IA et de la Robotique. Bref, une IA morale comme un capitalisme moral, qui n'existe déjà pas ? La maîtrise se veut aussi sociale avec de la formation continue pour adapter le marché du travail aux mutations de l'IA (on pourrait dans pas mal de cas y caser le numérique en général).

Côté **utile**, il était question de valorisation de la recherche fondamentale, et de la rendre plus transversale, d'encouragement à la création de champions européens de l'IA, d'orientation des investissements vers des applications sociétalement utiles, de création de cursus de formation sur l'IA[1885] et aussi de promotion de la diversité et de la place des femmes dans la recherche en IA[1886]. Le rapport préconisait la création de projets de recherche transversaux structurants et la création de champions européens. Mais cela ne se décrète malheureusement pas.

Et côté **démystification**, re-formation, mais dans le primaire et le secondaire, de la sensibilisation du grand public, la création d'un salon international de l'IA et de la robotique.

Rapport de la Mission Villani

Le Rapport de la Mission Villani était publié le 28 mars 2018 à l'occasion d'une belle conférence « AI for Humanity » organisée au Collège de France avec un beau panel d'intervenants français et étrangers comme Dennis Hassabis de DeepMind et bien évidemment, Yann Le Cun de Face-

---

[1884] Qui associe députés et sénateurs. Le mathématicien Cédric Villani a pris la présidence de cet office depuis son élection comme député LREM en juin 2017.

[1885] Je remarque que les cursus français ne donnent pas toujours lieu, comme dans les principaux cursus anglo-saxons, à la publication des supports de cours en ligne en libre accès. Voilà une belle source de progrès ! Comme à Stanford pour ce cours de traitement du langage : http://web.stanford.edu/class/cs224n/syllabus.html, le cours de Stanford de reconnaissance d'images d'Andrej Karpathy : http://cs231n.stanford.edu/syllabus.html, ou ce cours généraliste sur l'IA de l'Université d'Amsterdam dont les supports de présentation sont très bien faits : http://uvadlc.github.io. Heureusement, certaines universités font de beaux efforts, comme l'Université de Lyon 2 qui publie tous les supports de son cours de data mining et data science.

[1886] Un écueil qui est aussi soulevé dans le rapport AIReport de l'initiative **ArtificialIntelligenceNow** qui évoque le manque de diversité des chercheurs et développeurs de l'IA. Non seulement, les femmes sont sous-représentée mais aussi les minorités : « *Like all technologies before it, AI systems reflect the values of their creators, and there is hope that increased diversity among those developing, deploying, and maintaining AI systems may help create a future in which these technologies promote equality. Currently, however, women and minorities continue to be under-represented in the field of AI particularly, and in computer science overall* ».

book[1887]. Un œil exercé et un bon sens pratique ne peuvent en général qu'être déçus par ce genre de rapport et celui-ci ne faisait pas exception.

Le plan se focalisait surtout sur l'organisation de la recherche en France avec la création d'instituts 3IA voisins dans le principe des IRT et la revalorisation du salaire des chercheurs en IA[1888].

Suivaient l'adoption de l'IA par la puissance publique et les questions d'éthique. Il était très tourné sur les besoins du marché français. Le rapport manquait de pédagogie sur la dimension entrepreneuriale, n'insistant pas assez sur les spécificités des startups dans l'IA et les difficultés habituelles des entrepreneurs français à créer des produits.

Étaient tout aussi décevantes les parties sur les composants (introduction sur les chipsets neuromorphiques et proposition de créer un supercalculateur pour les chercheurs), sur la robotique (un plan franco-allemand vide comme l'éther), sur la santé (avec une vision trop franco-française des besoins) ou les véhicules autonomes (pas d'ambition clairement affichée). La partie formation était convenable, tout comme la volonté affichée d'attirer plus de femmes dans les filières de l'IA.

On peut se poser la question de la mesure de notre performance économique dans l'IA. La comparaison avec les USA et la Chine n'a pas de sens au vu de la taille de leurs économies respectives. Quels indicateurs (KPI) détermineraient que nous sommes « bons » dans l'IA en tant que pays ? Une production de valeur ajoutée à base d'IA supérieure au poids du PIB de la France dans le PIB mondial ? Un solde commercial positif dans le domaine ? Personne ne s'est encore aventuré à imaginer comment mesurer tout cela, à part le décompte de publications scientifiques et de leur influence chez les chercheurs. Cela peut aussi se mesurer en nombre de licornes, d'introductions en bourse (IPO), voir idéalement en chiffre d'affaires de startups de l'IA ou de recrutements dans l'IA[1889]. Mais comme dans tout plan public, le Rapport Villani ne propose aucun indicateur avec un horizon de temps ! Bref, on ne sait pas vraiment qualifier et quantifier les objectifs à atteindre ni donc les résultats.

Plan IA de l'État

À partir de septembre 2018, la coordination de la mise en œuvre du plan IA de l'État était prise en main par Bertrand Pailhès, ancien Directeur de Cabinet d'Axelle Lemaire à Bercy. Cela a conduit au déroulement d'une bonne partie des propositions contenues dans le rapport de Cédric Villani[1890]. L'une des actions clés a été le lancement des quatre 3IA, ces instituts de recherche appliquée en IA déjà cités. Il était suivi en mars 2020 par Renaud Vedel, jusqu'en novembre 2020.

[1887] J'avais abondamment commenté le contenu du rapport dans Ce que révèle le Rapport Villani, en mars 2018. À de rares exceptions près, l'accueil sur le rapport Villani était positif dans la presse française. Elle se félicitait que le sujet soit pris en compte et que l'éthique et les dimensions sociétales y soient mises en avant. Les réactions en creux comme la mienne n'étaient pas nombreuses. On peut citer IA et éthique: le contresens navrant de Cédric Villani de Philippe Silberzahn, Réaction au rapport Villani d'Alain Mueller qui portait sur le manque d'approche entrepreneuriale autour de la notion clé de product management, Intelligence artificielle, être suiveur ou devenir leader de Gilles Babinet, puis Et si l'intelligence artificielle n'était pas le problème du rapport Villani ? de Ariel Kyrou et Thierry Taboy qui posent la question sur la société que nous voulons créer avec l'IA.

[1888] La proposition consiste à doubler le salaire des chercheurs rémunérés par l'État débutants sachant qu'ils démarrent au CNRS à environ 36K€ bruts tandis qu'un ingénieur en IA chez Google démarrerait à 140K€. Mais il ne faut pas confondre chercheur et ingénieur. Le décalage entre les deux a toujours été élevé, quelle que soit la discipline. Le hic vient de ce que les GAFA recrutent aussi des chercheurs et au prix des ingénieurs, et cela a un effet inflationniste certain. Le secteur public ne peut pas suivre en Europe. Aux USA, il en va différemment car nombre d'universités qui emploient des chercheurs sont privées et ont plus de souplesse. Mais le salaire des chercheurs aux US n'est pas mirobolant non plus. Il n'a pas l'air de dépasser les $80K ce qui équivaut en intégrant le coût de la santé et de l'éducation à environ 50K€ en France. Dans les chiffres brandis par les uns et les autres sur les effectifs de R&D des grands acteurs du numérique, on confond souvent chercheurs et ingénieurs. Ces derniers sont la majorité dans les GAFAMI (les géants américains) et les NATU (les géants chinois) avec un rapport qui est souvent de 1 à 20 entre le nombre de chercheurs et d'ingénieurs/développeurs.

[1889] Voir La France, 7e pays recruteur de profils IA au niveau mondial par Véronique Arène, février 2019.

[1890] Voir Intelligence artificielle : la France dévoile son plan pour être leader mondial par Bastien L, novembre 2018 ainsi que Un an après le rapport Villani : où en est la stratégie nationale en IA ? par Remy Demichelis, mars 2019.

C'était complété par la création planifiée de dizaines de chaires et de dispositifs destinés à attirer les chercheurs et talents étrangers en IA. L'ANR dispose aussi d'une enveloppe de financement de projets de recherche en IA à hauteur de 100M€. Il y eu également le lancement du calculateur HPE au GENCI du CNRS à Saclay. Deux grands défis sur l'IA ont également été lancés par le **SGPI** (Secrétariat Général pour l'Investissement), sur le diagnostic médical et sur la sécurisation des systèmes d'information.

L'implémentation comprend aussi un appel à projets lancé en juillet 2019 par le gouvernement autour de la création création et/ou de l'enrichissement et/ou de l'exploitation de grandes bases de données mutualisées, centralisées ou distribuées, sectorielles ou multi-sectorielles. L'appel à projet était ouvert jusqu'en janvier 2020. Une bonne partie du plan de l'Etat vise à accélérer l'adoption de l'IA dans ses différentes missions, ce qui prend évidemment du temps.

Du côté économique, un plan était annoncé en juillet 2019 par Bercy. La sensibilisation des entreprises était au programme. Une partie de cette mission était pris en main par **Bpifrance** dans le cadre du financement de startups des deep techs, mais il y a fort à faire pour créer des leaders mondiaux[1891].

Le volet économique comprend trois grands défis avec 100M€ de financement (diagnostic médical, fiabilisation de l'IA et automatisation de la cybersécurité), 250M€ de « projets structurants » et huit Challenges d'innovation ouverte dotés en tout de 5M€, tous gérés par le SGPI (Secrétariat Général du Programme d'Investissements).

La gouvernance de ces projets est probablement complexe et je préfère ne pas en parler. L'Etat souhaite aussi encourager les initiatives de mutualisation des données entre acteurs privés via des Appels à Manifestation d'Intérêt gérés par la DGE.

Rapport de France Stratégie

Commandé par Muriel Pénicaud, ministre du Travail et Mounir Mahjoubi, secrétaire d'État chargé du numérique, le Rapport de France Stratégie sur l'Intelligence et le travail[1892] fait une analyse circonstanciée de l'impact de l'IA sur plusieurs secteurs : les transports, la banque et la santé. Le rapport est très mesuré et évite de sombrer dans le catastrophisme habituel. Pour les rapporteurs, l'impact de l'IA sur l'emploi dans ces différents sera à la fois lent et modéré.

Plan de l'Académie des Technologies

Quelques jours après la publication du rapport de la mission Villani, l'Académie des Technologies remettait le couvert avec son propre plan IA, coordonné par Yves Caseaux, l'un de ses membres et également DSI de Michelin.

Le plan se veut plus pratique et tourné vers les besoins des entreprises. Il souhaite aussi promouvoir la dimension européenne de l'IA tout comme sa dimension éthique[1893]. Il fait aussi de la pédagogie en décrivant les différents outils de l'IA destinés aux entreprises.

Il insiste sur le manque de compétences dans l'ingénierie des projets d'IA et propose la création de laboratoires d'essai et de certification de solutions d'IA par marché vertical tout comme la création d'un observatoire de l'Intelligence artificielle, placé sous le contrôle d'une agence européenne (à créer), pour suivre les bonnes pratiques de l'IA dans la société.

[1891] Voir Quelle politique industrielle pour l'intelligence artificielle en France ? par Sébastien Tran, août 2019, French Tech has made big strides but wants to go faster and further par Chris O'Brien, 2019 et L'intelligence artificielle au service des entreprises, stratégie nationale pour l'intelligence artificielle, presentation du volet économique, juillet 2019 (43 pages).

[1892] Voir Intelligence Artificielle et Travail, France Stratégie, mars 2018 (90 pages).

[1893] Voir le rapport Renouveau de l'intelligence artificielle et de l'apprentissage automatique et la présentation associée d'Yves Caseau, avril 2018.

Plan de la région île de France

En octobre 2018, la région île-de-France présentait son plan sur l'intelligence artificielle, conçu comme un rebond sur les propositions de la mission Villani. Il ambitionne de faire de la région le hub de l'IA en Europe en s'appuyant notamment sur l'écosystème d'enseignement supérieur et de la recherche publique et privés. Comme dans de nombreux plans, on y trouve à la fois une logique de dynamisation de l'offre notamment des startups et de la demande, en particulier des PME. Le plan s'articule sur quinze mesures dont la praticité est variable d'un cas à l'autre.

Un premier gros volet vise à dynamiser les usages de l'IA dans les PME. Les mesures ne sont pas évidentes à déployer : des packages de conseil et de PoC (projets pilotes) pour une centaine de PME, la création d'un inventaire de l'offre d'IA, la mutualisation des données industrielles, l'accès à une puissance de calcul souveraine (avec des GPU Nvidia), le projet Inriatech qui veut proposer une offre de recherche en IA à destination des PME et ETI, des formations BAC+2 pour les jeunes et les demandeurs d'emploi réalisées par Simplon.co pour Microsoft et la création d'un lycée de l'IA pour l'apprentissage du code dans l'IA (ce qui me semble un peu trop tôt, et pas associé à un Bac spécifique).

S'ensuit un soutient soutien financier au projet Digihall piloté par le CEA consistant à créer un écosystème d'innovation sur le plateau de Saclay (pas évident qu'il soit resté actif en 2020), le développement des coopérations internationales avec le Québec, la Bavière et la Corée du Sud et la communication sur l'excellence de l'IA francilienne. Enfin, et c'est le plus concret et réalisable, des challenges thématiques destinés aux startups dans l'oncologie et l'Hôpital du futur, dans l'industrie autour de l'apprentissage par transfert, les politiques publiques. Les lauréats du premier challenge IA étaient décernés le 15 octobre 2018 avec des prix conséquents étant de 100K€ (Panda, Smartify), 350K€ (LightOn) et 700K€ (Therapanacea).

Éthique

L'IA éthique est devenue une question sociétale clé pour faire accepter son déploiement. La thématique est étudiée depuis longtemps par les chercheurs comme Laurence Devillers ou Ravi Chatila en France[1894]. Le Rapport Villani s'était largement fait le relai du besoin de limiter les dérives du far-west de l'IA en mettant en avant cette notion d'IA éthique.

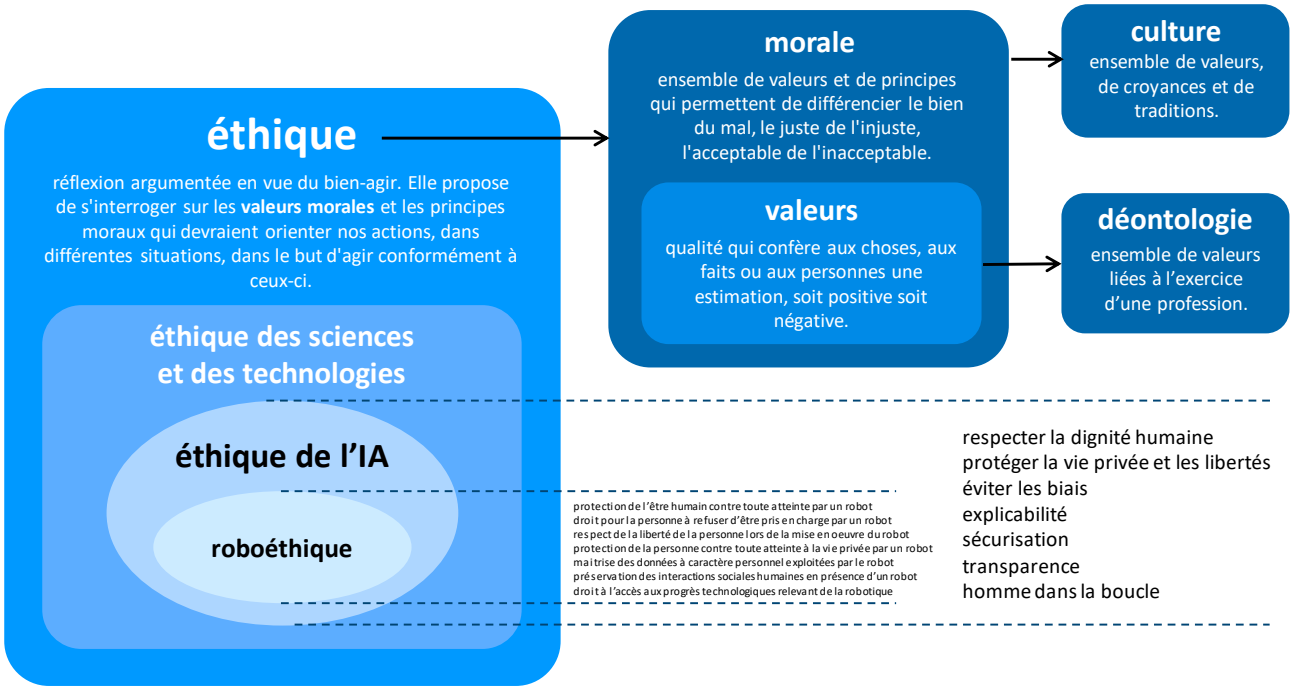
Ce concept aurait très bien pu être accolé aux logiciels et à nombre de technologies depuis qu'elles existent mais l'approche probabiliste du machine learning et du deep learning a mis son grain de sel dans l'équation pour justifier de s'intéresser de plus près aux conditions de la mise en œuvre de l'IA[1895]. C'est anodin lorsque le machine learning permet à Netflix de nous recommander une série B correspondant à nos goûts. Ça l'est moins pour une IA qui gère un diagnostic médical, pilote un véhicule autonome ou décide de nous octroyer un prêt bancaire. La question est encore plus épineuse pour les robots qui interagissent avec les humains, surtout les robots humanoïdes sans compter les systèmes d'armes (éventuellement) autonomes.

L'éthique de l'IA est aussi étudiée pour éviter les risques les plus souvent associés aux usages de l'IA[1896].

[1894] Voir Questionnements éthiques sur la robotique et l'intelligence artificielle par Raja Chatila, 2017 (46 slides) et La recherche d'une éthique de l'IA par Marina Ehrhart, octobre 2019.

[1895] Voir IA et éthique ? Calmons nous par Jérome Fortias, janvier 2020, qui explique bien pourquoi l'éthique de l'IA est une focalisation sur un type de technologie qui en concerne en fait bien d'autres. Point de vue que l'on retrouve dans The Importance of Ethics in Artificial Intelligence par Ferry Hoes, décembre 2019.

[1896] Le MIT a listé six risques de l'IA : les véhicules autonomes, les manipulations spolitiques, les algorithmes tueurs, la reconnaissance faciale, les deep fakes et la discrimination algorithmique. Curieusement, aucun risque d'évoqué dans le domaine de la santé ! Voir Les 6 dangers de l'IA selon le MIT par Adrienne Rey, février 2019.



La tendance est donc de faire de l'IA éthique « by design » comme on fait de la « privacy by design », les deux notions étant d'ailleurs étroitement liés. L'IA éthique by design requiert des objectifs de moyens pour les créations d'applications exploitant de l'IA. La question clé pour les gouvernants est de savoir si l'on met en place une réglementation a priori ou a posteriori de la création des innovations.

Comment cependant distinguer éthique, éthique de l'IA, morale, valeurs et déontologie ? Voici un petit schéma de mon cru qui permet d'appréhender les liens entre ces concepts philosophiques.

- Une **éthique** est réflexion argumentée en vue du bien-agir. Elle propose de s'interroger sur les valeurs morales et les principes moraux qui devraient orienter nos actions, dans différentes situations, dans le but d'agir conformément à ceux-ci. Elle s'applique en général ou à des domaines spécifiques comme les sciences et les technologies et à l'intelligence artificielle qui en est un sous-ensemble. Une éthique est donc surtout un cadre de travail et un ensemble de principes généraux de gouvernance d'un domaine donné. Elle sera généralement décomposée en objectifs de moyens (« mettre l'Homme dans la boucle de décision ») et de résultats (« créer des IA sans biais et/ou explicables »)[1897]. L'éthique de l'IA est souvent définie en creux pour éviter l'émergence de travers comme des atteintes à la vie privée ou à la dignité humaine, notamment par la création de biais. Mais on ne parle par d'éthique du train, de l'avion ou de la cocotte-minute.
- Une **morale** est ensemble de valeurs et de principes qui permettent de différencier le bien du mal, le juste de l'injuste, l'acceptable de l'inacceptable. Elle est souvent relative à une culture donnée. Ce n'est pas un absolu.
- Une **valeur** est une qualité qui confère aux choses, aux faits ou aux personnes une estimation, soit positive soit négative. Là encore, elle est relative à une culture. Ce qui est bien chez les Chinois ne l'est pas forcément en Occident et réciproquement.
- Une **déontologie** est un ensemble de valeurs liées à l'exercice d'une profession, comme par exemple, celle des médecins. Mais elle pourrait être créée pour des développeurs, des data-scientists et autres spécialistes de la création d'IA.

[1897] L'éthique téléologique s'intéresse uniquement au résultat…

- Une **culture** est ensemble de valeurs, de croyances et de traditions souvent associée à un pays, une religion ou une population.

En pratique l'éthique de l'IA couvre plusieurs aspects que les entreprises doivent intégrer dans leurs réflexions, organisation, création de solutions intégrant de l'IA et sur leur communication[1898] :

- Cela comprend d'une manière générale le principe de bienfaisance consistant à **faire le bien** et celui de non-malfaisance consistant à **ne pas faire le mal**. S'y ajoute le principe de **justice**. Mais ce n'est pas spécifique à l'IA.
- Respecter la **dignité humaine** qui est une déclaration générale de bonne intention mais qui n'est pas spécifique à l'IA. Elle s'applique à n'importe quelle activité économique ou politique.
- Protéger la **vie privée et les libertés**, qui s'applique aux usages du numérique en général, mais dans lesquels l'IA joue un rôle de plus en plus grand[1899]. Sachant que ces notions varient d'un pays à l'autre et notamment de leur système culturel et politique.
- Eviter les **biais de l'IA** générés par des algorithmes mal conçus ou des jeux de données d'entraînement inadéquats.
- Créer des **IA explicables et auditables**, surtout pour les applications les plus critiques dans la santé, la police, la justice ou les transports autonomes.
- Faire preuve de **transparence** dans la création des IA, qui rejoint le point de l'auditabilité et de l'explicabilité. Il est souvent étroitement associé à l'exigence d'utilisation de logiciels libres[1900].
- Créer des **systèmes sécurisés** qui relève d'objectifs de moyens permettant d'assurer les points précédents.
- Intégrer les **humains dans la boucle** des usages de l'IA, notamment pour les décisions les plus critiques. On peut y intégrer l'exigence de diversité des équipes qui créeent des IA, notamment dans l'équilibre entre les hommes et les femmes qui est plutôt exécrable en général avec moins de 10% de femmes dans les équipes techniques de création d'IA[1901].
- Il faudrait aussi y ajouter la notion de **conscience sur l'impact à long terme** de l'IA et surtout en termes environnementaux. La consommation d'IA à outrance coûte cher en énergie primaire et donc, dans la plupart des pays, en énergies fossiles émettrices de CO2 dans l'atmosphère. L'empreinte énergétique de l'IA est noyée dans la grande masse de la consommation d'énergie liée aux usages du numérique.

[1898] Voir Pour y voir plus clair parmi les très nombreux documents qui traitent d'éthique+IA par Aymeric Poulain Maubant, février 2020 ainsi que Top 12 AI Ethics Research Papers Introduced In 2019 par Mariya Yao, 2019.

[1899] Voir L'IA du Quotidien peut elle être Éthique ? par Philippe Besse, Céline Castets-Renard, Aurélien Garivier et Jean-Michel Loubes, décembre 2018 (28 pages) qui étudie l'impact du RGPD sur l'éthique de l'IA et surtout des modèles de machine learning. Il décrit par quelques exemples les différents biais statistiques qui peuvent être introduits par le machine learning. On y trouve aussi des notions associées à la differential privacy consistant à répartir des erreurs par sous-classe de population pour éviter qu'elles soit concentrées sur une classe, comme pour la détection de récidive dans des applications judiciaires US qui surpondéreraient la population noire. Voir notamment How We Analyzed the COMPAS Recidivism Algorithm — ProPublica par Jeff Larson,Julia Angwin, Lauren Kirchner et Surya Mattu, 2016.

[1900] En pratique, de nombreuses IA ne sont pas construites sur des logiciels libres de bout en bout. Seuls les outils de développement le sont d'un point de vue pratique. Allez récupérer le code source de Google Search pour voir !

[1901] Voir Intelligence Artificielle et Robotique: Quelle éthique ? par Raja Chatila, février 2019 (17 slides) qui considère que les machines ne prennent pas de décisions éthiques. Elles peuvent par contre prendre des décisions qui ont des conséquences éthiques. C'est un point de vue qui ne fait pas l'unanimité. En fait, l'éthique est un cadre. Elle repose sur des valeurs qui peuvent être parfois encodées dans des systèmes. Pour Raja Chatila, les algorithmes sont conçus par les humains et les machines qui les exécutent n'ont pas une compréhension des fondements de leurs décisions et de leurs actions. La dignité humaine n'est pas explicitement décrite et ne peut être apprise à une machine.

- Le déploiement d'une **pédagogie** sérieuse sur ce qu'est et fait l'IA vis-à-vis du grand public. Comme tout sujet nouveau, il est important que les citoyens soient éclairés sur le fonctionnement et les impacts de l'IA.
- Des principes éthiques sont aussi créés pour les systèmes autonomes et intelligents comme avec l'**IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems**. Comme pour l'IA en général, ces principes visent à respecter les droits humains, à l'amélioration du bien-être humain, à l'aptitude à l'objectif (le système doit être conçu de manière à accomplir la tâche attendue et être prévisible), à la responsabilité (elle demeure celle des humains qui sont derrière les systèmes), à anticiper et prévenir les més-usages et conséquences inattendues.

L'une des spécialistes de l'éthique de l'IA en France est **Laurence Devillers**, chercheuse au CNRS et auteur du livre « Des robots et des hommes ». Elle met en avant divers risques à venir : le risque émotionnel lié à un attachement trop fort aux robots[1902] ou au contraire un rejet du robot et celui d'une société de plus en plus paresseuse, où l'on céderait trop souvent son libre arbitre aux machines au point de permettre un contrôle de sa vie pouvant devenir totalitaire.

Cela pourrait créer un gouffre encore plus béant qu'aujourd'hui entre ceux qui connaissent les entrailles des systèmes et la majorité des utilisateurs. La question se poste pour les robots ayant une apparence physique ainsi que pour les logiciels sans incarnation physique autre que du texte, des avatars en 2D ou de la voix. Nombre de ces questions se posent même sans IA, comme dans les usages des mobiles ou dans les jeux vidéo.

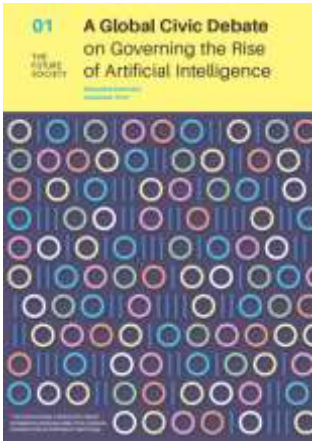
Quand Snapchat créé l'addiction chez les adolescents en réduisant leur popularité sur le réseau social s'ils ne publient rien, il créé un système pavlovien asservissant. Et pourtant, sans IA. Laurence Devillers propose la mise en place de règles de bonne conduite pour que les IA et les robots soient éthiques dès leur conception, pour éduquer les utilisateurs et surtout les plus jeunes, la création d'outils de vérification de ces règles.

Cela irait jusqu'à la création d'un cadre légal et pénalisant en cas de non-respect de ces règles, le tout étant encadré par des comités éthiques indépendants. Sa praticité ne serait pas triviale.

L'un des défis est de concilier les modèles économiques des acteurs du marché avec ces éléments d'éthique, notamment lorsque sont mis en œuvre des modèles publicitaires et addictifs. Même si le respect de l'éthique ne devrait pas être dépendant des modèles économiques de ces acteurs.

[1902] Voir Robot : une étude prouve qu'on ressent des sentiments pour une IA, de Pierrick Labbé, août 2018.

Le souci de l'éthique de l'IA n'est pas spécifique à la France. C'est une préoccupation mondiale, ou tout du moins occidentale. Nombre d'associations ou consortiums d'origines américaines ont été créées dans cette mouvance comme **Open AI**[1903], **The Future Society**, **Partnerships on AI** (créé par les GAFAMI) ou encore l'**AI Now Institute**, de l'Université de New York et qui comprend notamment des chercheurs de Microsoft, qui publiait un rapport en décembre 2018 demandant une régulation par les états des usages de l'IA[1904].



La déclaration de Montréal de mars 2018 est une initiative lancée par l'**Université de Montréal**. Elle résulte d'un processus de « co-construction » qui faisait suite au Forum on the Socially Responsible Development of Artificial Intelligence, de novembre 2017.

The Future Society[1905] publiait en septembre 2018 A global civil debate on governing the rise of artificial intelligence (13 pages), une proposition de gouvernance de l'IA. En mai 2019, le **World Economic Forum** lançait un groupe de travail d'acteurs privés du secteur de l'IA (Microsoft, Uber, des Chinois, etc) se donnant comme objectif de restaurer la confiance dans l'IA avec comme objectif l'organisation d'un sommet en avril 2020[1906].

Dans le lot, il y a aussi le **Future of Life Institute** créé par Max Tegmark, l'auteur de Life 3.0 qui décrit un futur plein d'IA et les risques associés. Y sont définis les principes d'**Asilomar**, du nom d'une conférence organisée en janvier 2017 en Californie avec une centaine de participants représentant le gotha occidental de l'IA[1907].

Les principes recouvrent : la robustesse des IA, le financement équilibré d'IA et de l'analyse de leur impact, un impact social équilibré de l'IA, des IA auditables (ou explicables), un alignement des IA sur les valeurs humaines, le droit à la vie privée, l'évitement d'une course aux armes autonomes et un développement d'éventuelles AGI qui devrait être géré avec une infinie précaution surtout pour des versions s'améliorant de manière récursive. Bref, on en est à définir l'éthique d'un objet qui est pour l'instant mythique, l'AGI.

Les détracteurs de ces démarches rappellent que les notions d'éthique sont très variables selon les pays, cultures et religions. En Chine par exemple, l'harmonie d'ensemble de la société est plus importante que les libertés individuelles. D'où le système politique qui y sévit et les libertés limitées, comme dans l'usage sans vergogne de la vidéo-surveillance ou la notation des citoyens dans leurs pratiques sociales.

Il existe une autre notion d'éthique, plus générale, qui consiste à utiliser l'IA pour des solutions qui sortent de la sphère économique traditionnelle et visent le bien social, notamment en faveur des populations les plus défavorisées.

[1903] Voir OpenAI Director Shivon Zilis explains why AI requires oversight now de Darrell Etherington, octobre 2017.

[1904] Voir AI Now Report 2018 (62 pages) et AI desperately needs regulation and public accountability, experts say par Devin Coldewey, décembre 2018, qui y fait référence.

[1905] Voir leur site http://www.thefuturesociety.org/. C'est une ONG avec quelque Français dans ses advisors : Brice Lalonde, Hubert Védrine advisor et Nohza Boujemaa d'Inria.

[1906] Voir World Economic Forum Inaugurates Global Councils to Restore Trust in Technology par Amanda Russo du WEF, mai 2019.

[1907] Voir The Asilomar AI Principles et la conférence Beneficial AI 2017.

C'est la « tech for good » et l'entrepreneuriat social et solidaire (ESS) qui ont sa déclinaison « **AI for good** », correspondant même à une initiative du même nom sous l'égide des Nations Unies[1908].

A noter que l'Union Européenne a créé le 14 juin 2018 un comité consultatif de 52 experts, le **High Level Group on Artificial Intelligence**. Il doit servir d'éclaireur de l'Union Européenne pour la sauvegarde des droits humains et, pourquoi pas, proposer des dispositions réglementaires de l'IA.

La France a fait de même en décembre 2019 en créant un comité d'éthique sur l'IA de 27 personnes qui comprend surtout des chercheurs. Ce comité dépend d'un autre comité, le Comité Consultatif National d'Éthique. Il doit se focaliser sur les agents conversationnels, les véhicules autonomes et le diagnostic médical et devait remettre un premier rapport début 2021[1909].

Du côté des entreprises, l'IA va faire rapidement partie des obligations de leur Responsabilité Sociale et Environnementale. Elles devront afficher la manière dont elles intègrent l'éthique de l'IA dans les solutions qu'elles exploitent aussi bien à destination de leurs salariés que de leurs clients.

Divers cadres d'éthique de l'IA ont été créés pour les entreprises[1910], notamment provenant de grandes sociétés de conseil qui se sont emparées du sujet. On en trouve notamment chez **PwC** qui a créé un kit d'éthique de l'IA[1911] ainsi que chez **Capgemini**[1912] ou chez l'association internationale **ITechLaw**[1913].

Des entreprises françaises et des filiales françaises de sociétés de la tech ont créé l'association Impact IA qui vise à fédérer les efforts de création d'IA éthique, **Impact IA**. La présidence était assurée en 2019 par Laurence Laffont, de Microsoft France, qui est partie en 2020 chez Google[1914].

Vie privée

La notion de vie privée ramène immédiatement à l'application du **RGPD** en Europe, cette réglementation dont, en tant qu'utilisateurs, nous pouvons apprécier au quotidien l'absurdité avec ces validations pavloviennes de cookies qui nous sont imposées par tous les sites web que nous visitons. La perte de temps associée est significative.

Derrière cet agacement se cachent des procédures que les entreprises doivent adopter pour préserver la confidentialité et la sécurité les données collectées sur les utilisateurs. Ce sont des objectifs de moyens clairement documentés. Cela comprend ainsi le consentement à recevoir des spams qui ne peut plus être précoché, l'obligation de suppression des données sur demande, et automatiquement au bout de trois ans d'inactivité et l'obligation de fournir un extrait des données sur demande.

[1908] Voir le AI for Good Global Summit organisé par l'ITU à Genève en mai 2018. Une déclinaison française était lancée à l'Hôtel de Lassay sous l'égide de France is AI, France Digitale, et en présence de Cédric Villani le 3 avril 2018. Le principal sponsor de l'événement était … Facebook, en pleine affaire Cambridge Analytica. Des associations liées à la tech for good étaient aussi de la partie : la Croix-Rouge française, Bibliothèques sans frontières, Green Cross tout comme la startup DataforGood qui est notamment associée à la startup Bayes Impact de Paul Duan.

[1909] Voir Création du Comité Pilote d'Éthique du Numérique, décembre 2019.

[1910] Voir Seven very simple principles for designing more ethical AI par Frida Polli, août 2019, qui décrit la mise en oeuvre d'une IA éthique dans les entreprises en sept points : la compréhension des mécanismes de collecte de données, le respect de la vie privée, l'évitement de biais des données, l'auditabilité des systèmes, la transparence des IA (qui revient au point précédent), l'usage d'outils et méthodes open source et l'appel à des garde-fous éthiques externes.

[1911] Voir le PwC : responsible AI toolkit, 2019 (20 pages). Gouvernance, éthique et régulation, interprétabilité et explicabilité, robustesse et sécurité, biais et équité. Et Intelligence artificielle une révolution en marche, mars 2019 (60 pages).

[1912] Voir Capgemini : why addessing ethical questions in AI will benefit organizations, 2019 (34 pages).

[1913] Voir Responsible AI: A Global Policy Framework, 2019 (308 pages) par l'association ITechLaw (down).

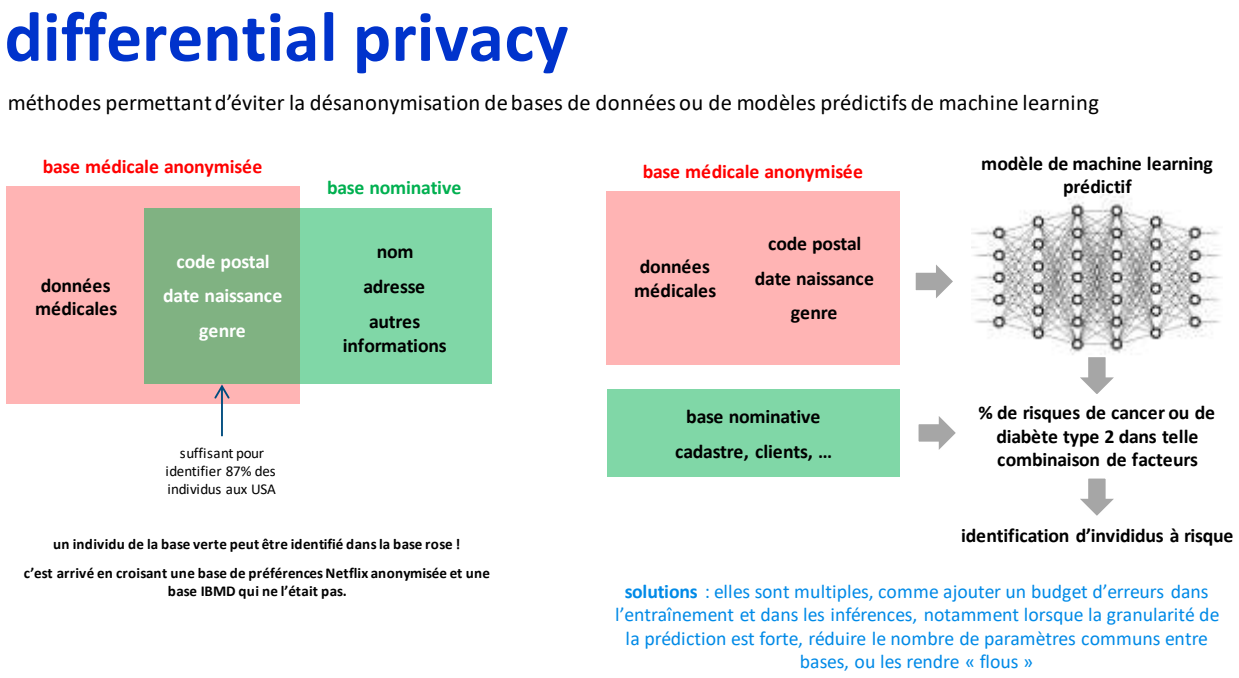
[1914] Voir Impact IA : un engagement collectif pour un usage responsable de l'intelligence artificielle, 2018 (21 pages).

Qu'en est-il de ces données une fois qu'elles passent à la moulinette de l'IA ? Lorsqu'elles sont exploitées par des briques de machine learning et de deep learning, elles perdent généralement leur caractère nominatif. Il n'y a pas plus anonyme qu'un gros réseau de neurones entraîné avec des paramètres décrivant le comportement d'un grand nombre de clients.

Ces réseaux de neurones servent ensuite à faire des classifications et des prévisions sur le comportement de nouveaux clients ou de clients existants, basés sur le comportement collectif.

La préservation de la vie privée n'est pas garantie avec des données anonymisées dès lors qu'elles intègrent plus d'une dizaine de paramètres[1915]. Notre signature personnelle est en effet unique dès que l'on combine une masse critique de ces paramètres. Si certains de ces paramètres sont communs avec ceux d'une autre base, elle non anonymisée, alors, on peut faire des recoupements et désanonymiser une base de donnée anonymisée[1916].

Le principe de la « *differential privacy* » doit permettre d'éviter qu'une entrée de donnée anonymisée d'un utilisateur ait un impact significatif sur les résultats de l'algorithme, permettant son identification éventuelle en croisant le résultat avec une autre base de donnée. Par exemple, si on cherchait à obtenir une statistique sur un échantillon d'utilisateurs qui est trop réduit, on pourrait obtenir l'information associée sur un utilisateur particulier de cet échantillon.



Les techniques couramment utilisées consistent à ajouter du bruit dans les données anonymisées qui alimentent le modèle. Du bruit est surtout ajouté en sortie pour éviter que le modèle puisse être utilisé pour identifier des utilisateurs particuliers.

[1915] Voir Artificial intelligence and privacy Report, du norvégien Datatilsynet, janvier 2018 (30 pages) qui fournit une approche européenne sur la manière de tenir compte du RGPD dans sa stratégie de données et d'IA ainsi que Artificial intelligence and privacy Issues paper, Office of the Victorian Commissioner, juin 2018 (15 pages).

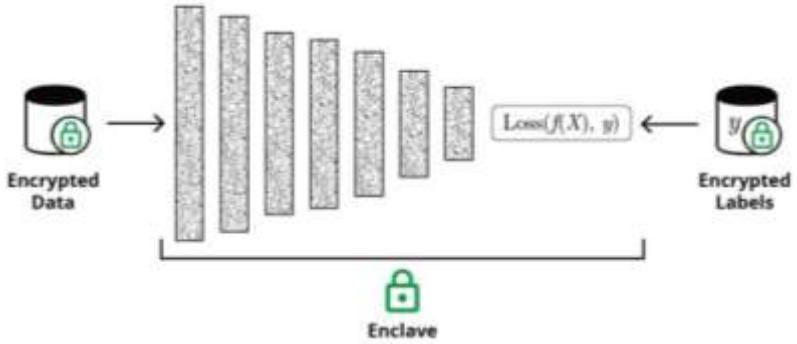
[1916] La désanonymisation des données privées et les risques associés sont les thèmes de recherche de Sonia Ben Mokhtar, du CNRS LIRIS à l'Insa de Lyon. Voir son site https://sites.google.com/site/soniabm. Dans Artificial Intelligence as as Digital Privacy Protector, 2017 (19 pages) d'Andrea Scripa Els de Harvard explique comment l'IA peut améliorer la vie privée et contourner ces problèmes. Avec notamment la notion de « Differencial Privacy ». Le processus de réidentification d'utilisateurs est bien décrit dans la présentation Big Data, Artificial Intelligence and Privacy de Stephen Kai-yi Wong (responsable de protection de la vie privée à Hong Kong), 2017 (34 slides). Dans la même lignée, Latanya Sweeney est connue pour ses travaux sur la collecte d'informations privées par les applications mobiles. Voir Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps, octobre 2015. Elle est aussi la fondatrice du Privacy Data Lab à Harvard.

C'est particulièrement important pour les applications dans la santé, les assurances ou lors de recensements[1917]. Apple et Microsoft sont de grands promoteurs de la *differential privacy*.

On peut aussi alimenter des modèles de machine learning avec des données chiffrées. Les méthodes de chiffrement dites homomorphes permettent de le faire en théorie. Un tel chiffrement permet de réaliser des opérations mathématiques comme des additions et multiplications sur des données chiffrées et de déchiffrer ensuite le résultat.

Cela permet d'alimenter des algorithmes avec des données chiffrées, ce qui peut être utile pour des services réalisés dans le cloud. Une entreprise ou organisation peut les entraîner les modèles avec des données chiffrées et exécuter les modèles entraînés avec des entrées et des sorties chiffrées. Les opérations de chiffrement et déchiffrement ont lieu chez le client. L'opérateur du cloud ne peut ainsi pas accéder aux données et exploiter le modèle entraîné.

C'est un domaine en plein devenir. Les premiers systèmes entièrement homomorphes sont apparus en 2009[1918]. Seuls quelques types de machine learning et deep learning peuvent être alimentés avec des données chiffrées de manière homomorphe, comme les Crypto-Nets.



Le véritable risque, associé au fameux biais des données d'entraînement, est que votre cas particulier ne soit pas représenté en quantité suffisante d'un point de vue probabiliste dans ces données d'entraînement ou qu'au contraire il soit surreprésenté. On peut cependant aussi entraîner des IA avec uniquement vos données personnelles si elles sont en quantité suffisante, par exemple, pour vous faire des recommandations de produits, contenus et services. C'est aussi utilisé dans certaines applications de formations en ligne personnalisées.

Les données des utilisateurs sont généralement anonymisées avant d'entraîner des IA probabilistes. Ces bases intermédiaires ne sont pas directement utilisées. Ce sont des sous-produits d'un processus au bout duquel on exploite des modèles d'IA probabilistes entraînés. Les données amont de l'IA sont donc plus sensibles que les modèles entraînés avec.

Parmi les spécialistes du chiffrage homomorphe, on compte notamment **Cosmian** (2018, France, 1,4M€), **CryptoExperts** (2008, France), **Enveil** (2016, USA, $4M), **Kindite** (2017, Israël, $4M), **Inpher** (2014, USA, $14M), **Nucypher** (2015, USA, $5,2M) et **Zama** (2020, France)[1919].

[1917] La méthode a été créée par Cynthia Dwork en 2006. Voir Differential Privacy A Primer for a Non-technical Audience, février 2018 (41 pages), et la présentation associée, 2017 (45 slides), Deep Learning with Differential Privacy, 2016 (14 pages), Differential privacy and machine learning (127 slides), Differential privacy: an introduction for statistical agencies, par Dr. Hector Page & AL, décembre 2018 (53 pages) et Privacy-preserving machine learning and differential privacy par Antti Honkela, décembre 2018 (21 slides). Pour l'application de differential privacy lors de recensements, voir Differential Privacy in the Real World: The 2018 End-to-End Census Test, par John M. Abowd, février 2019 (33 slides).

[1918] Voir Crypto-Nets: Neural Networks over Encrypted Data, 2015 (9 pages), la présentation An Introduction to Homomorphic Encryption for Statistics and Machine Learning de Louis Aster, 2015 (68 slides) et Deep Learning on Private Data de Sadegh Riazi, Bita Darvish Rouhani et Farinaz Koushanfar (9 pages).

[1919] Zama a été lancée par Rand Hindi, ancien fondateur de Snips, acquise par Sonos fin 2019, en s'appuyant sur les travaux de Pascal Paillier qui fait aussi partie de la société CryptoExperts. Voir le cryptosystème de Pallier.

Comme le rappelle ce visuel *ci-dessous* issu de la **Quadrature du Net**, les grands acteurs de l'Internet savent beaucoup de choses sur la vie des utilisateurs. Heureusement, à moins d'avoir toute sa vie numérique chez Google ou chez Apple, nos données privées sont généralement réparties dans plusieurs de ces systèmes et ils ne se parlent pas (encore). Mais qu'en serait-il dans une ville intelligente très intégrée[1920] ?

Une question clé subsiste : comment faire d'un respect scrupuleux de l'éthique et de la vie un avantage marketing capable de renverser la situation ? Dans des marchés existants, cela semble difficile tellement les inerties sont grandes pour faire changer les habitudes tant des consommateurs que des entreprises, surtout si elles sont interdépendantes d'un réseau et d'un écosystème dense de solutions tierces parties.

C'est ce qui explique la stabilité de l'empire Microsoft dans les serveurs et les postes de travail depuis plus de 35 ans, celle d'Apple dans les smartphones depuis 11 ans, celle de Google dans les moteurs de recherche, depuis maintenant 20 ans et de Facebook dans les réseaux sociaux.

Des acteurs comme le moteur de recherche du Français **Qwant** (2011, France) mettent en avant le respect de la vie privée dans leurs bénéfices clés. Ils ne conservent en effet aucune donnée des utilisateurs pour rendre leurs recherches pertinentes. C'est particulièrement intéressant dans la version de leur moteur de recherche qui est destinée aux enfants.

C'est un avantage concurrentiel en Europe où le souci de la vie privée est le plus élevé, particulièrement en Allemagne et pour des raisons historiques liées (Stasi, Nazis). Certains utilisateurs et organisations – surtout dans le secteur public – abandonnent explicitement Google Search pour cette raison.

Le dilemme est que cela ne permet pas de cibler les utilisateurs de manière aussi pertinente qu'en conservant leurs données. D'où un revenu par utilisateur qui est structurellement plus faible. Cela rend difficile une concurrence face à Google d'un point de vue macro-économique.

Non seulement, ils bénéficient d'économies d'échelle plus grandes mais également d'un bien plus gros revenu par utilisateur (ARPU), nonobstant le fait que l'ARPU de Qwant n'est pas publié. Les solutions pour contourner cela ? Faire de ce service un « utility » régulé voir financé par les pouvoirs publics et veiller à garantir la concurrence de ce marché qui ne devrait pas tolérer qu'un seul acteur s'arroge plus de 90% du marché en Europe.

Il est difficile de changer la donne d'un marché établi avec uniquement cette fonctionnalité du respect de la vie privée. Le coût de l'éducation du marché est très élevé[1921]. Les inerties d'écosystèmes pèsent tout autant. Il vaut mieux être les premiers à générer des ruptures technologiques ou d'usages pour transformer radicalement un marché. Donc, pour de nouveaux usages, l'approche éthique/vie privée pourrait être plus porteuse, comme dans la santé, la ville intelligente ou les transports.

[1920] Voir Utopian Vision, Dystopian Reality, 2017 (25 pages) de Privacy International qui s'inquiète à juste titre des libertés individuelle dans une ville hyper-connectée.

[1921] Voir Intelligence artificielle : l'éthique est-elle soluble dans l'économie ? de Christophe Auffray, décembre 2017.

Il n'empêche que l'adoption de pratiques éthiques dans le numérique, avec ou sans IA, est incontournable pour les startups comme pour les entreprises établies. C'est un impératif sociétal qu'il sera de plus en plus difficile de contourner, soit du fait de la régulation, soit du fait de la réaction de la société en général comme l'a montré l'affaire Cambridge Analytica vis-à-vis de Facebook.

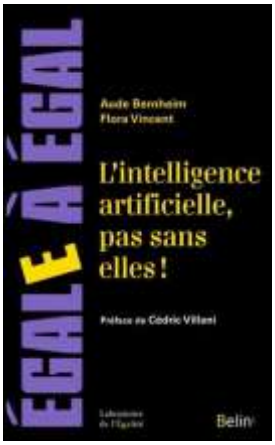
Diversité

L'une des questions clés de l'éthique est l'évitement du biais des algorithmes et des données dans l'entraînement des IA probabilistes, à base de machine learning et de deep learning.

L'un des moyens d'y parvenir consiste à rendre les équipes qui la créent plus diverses, à commencer par l'équilibre des genres qui est très mauvais à peu près partout dans le monde.

Que cela repose ou pas sur des stéréotypes, les hommes et les femmes ont en moyenne une approche différente des besoins applicatifs. Les solutions créées principalement par des hommes présentent des biais, ne serait-ce que thématiques[1922]. Elles ne peuvent pas prétendre incarner l'ensemble des besoins de la société. C'est l'une des raisons pour lesquelles il faut plus de diversité dans les équipes de chercheurs, d'ingénieurs et de cofondateurs de startups.

Aude Bernheim et Flora Vincent expliquent bien ce phénomène dans « L'intelligence artificielle, pas sans elles », paru en mars 2019 (91 pages)[1923].



Diverses initiatives visent à encourager plus de jeunes filles à s'orienter vers les métiers scientifiques et en particulier ceux du numérique, comme la Fondation **Femmes@Numérique**, lancée fin juin 2018 à la Cité des Sciences et de l'Industrie de la Villette en présence de quatre ministres. Elle servira à financer les actions sur le terrain de plus d'une quarantaine d'associations dont par exemple **Talents du Numérique**, ex Pascaline. Et il y a fort à faire car on assiste même à une régression avec de moins en moins de femmes qui sont attirées par ces carrières scientifiques et techniques.

Aucune baguette magique ne va résoudre ce problème instantanément. C'est l'affaire de presque une génération. Cela passe par la modification du rôle des filles et des femmes dans les médias, dans l'éducation, dans les jouets, par la valorisation des domaines d'applications du numérique et de l'IA dans des secteurs tels que la santé et l'environnement, par la valorisation de l'utilité sociale des solutions tout comme par celle de modèles féminins inspirants[1924].

On peut aussi saluer les initiatives **Women in Machine Learning**[1925] (USA) ainsi que **Women in AI**, qui a une branche en France.

Pédagogie

Un dernier point mériterait d'être traité côté éthique : la pédagogie. On ne peut avoir de débats sains sur l'éthique de l'IA que si on comprend les réalités technologiques et d'usages de l'IA.

[1922] Nous en avons un cas extrême au Japon avec **Gatebox** (Japon) et sa copine virtuelle, qui permet de rompre la solitude du jeune homme célibataire. Cela s'explique par la complexité des relations sociales intersexes chez les jeunes adultes au Japon (vidéo).

[1923] Voir « Le manque de femmes dans l'intelligence artificielle accroît le risque de biais sexistes », une interview de Aude Bernheim et Flora Vincent par Claire Legros, mars 2019 ainsi que L'IA est-elle sexiste, raciste, homophobe et transphobe ? par Irina Coyssi, 2019.

[1924] Voir par exemple Quand l'IA se code au féminin, février 2018 qui fait un inventaire de *role models* féminins qui ont joué un rôle dans l'histoire du logiciel, 100 women in AI ethics de Mia Dand en octobre 2018 et bien entendu l'initiative Quelques Femmes du Numériques !, projet photo lancé en 2012 et devenu en 2016 une véritable association qui contribue à la présentation de jeunes role models dans les lycées et établissements d'enseignemetn supérieurs (depuis 2017 dans les Pays de la Loire et bientôt en Rhône-Alpes).

[1925] Au sujet de Women in Machine Learning, voir The Women Changing The Face Of AI, août 2016.

L'IA est très diverse à ces deux niveaux comme nous avons pu le découvrir dans cet ebook. Or cela manque vraiment. L'IA est fantasmagorique. Sa perception et sa compréhension sont trop l'affaire de futurologues et singularistes qui confondent la réalité scientifique d'aujourd'hui et de demain avec les fantasmes de la science-fiction. Trois audiences sont clés pour faire avancer les débats : les politiques, les entreprises et le grand public. Ils ont besoins d'un tronc commun de connaissances et de messages différents. Cela explique notamment quelques initiatives qui veulent faire rentrer le numérique, le code et l'IA dans l'enseignement secondaire. Les approches mêlant institutions et société civiles peuvent marquer les esprits comme cette initiative du collectif et cluster **NaonedIA** à Nantes en juin 2018 qui publiait le Manifeste de Nantes, reprenant la déclaration de Montréal[1926].

L'influence médiatique en matière d'IA comme en politique est dominée par des personnalités extrêmes. Nous avons en premier lieu le trouble-fête, le poil à gratter et l'épouvantail en chef de l'IA qu'est **Laurent Alexandre**.

Puis des évangélistes tels que **Stéphane Mallard**[1927] et enfin, des personnalités modérées et écoutées comme **Yann Le Cun**, **Luc Julia**[1928] et plus récemment **Alexandre Templier**[1929]. Nous avons besoin d'un équilibre entre ces différents points de vue.

Nous devons faire preuve de pédagogie sur les réalités technologiques et des usages de l'IA. Une explication est nécessaire. Il faut ouvrir un peu la boîte noire. Tout comme le moteur à explosion n'explose pas au sens littéral du terme, l'IA est moins dangereuse que ses Cassandre le laissent croire. Aidons le public à décoder la propagande d'études de cas enjolivées que nous avons pu évoquer dans ce document. Dénonçons l'escroquerie des robots soi-disant autonomes de Boston Dynamics qui ne le sont pas (encore) dans la réalité !

Évitons les généralisations à outrance comme l'interprétation des images médicales par le deep learning qui est extrapolée à tous les actes de santé. N'oublions pas la composante humaine de tous les métiers voués à disparaître selon les prévisions les plus sombres. Il y a quelqus années, Laurent Alexandre secouait les neurones du monde politique en exagérant le trait, en crédibilisant parfois les thèses de singularistes, ou celles de la connexion hommes machine de Neuralink qu'il sait improbables[1930].

Cela fait réfléchir sur de graves problèmes de société qui sont bien réels, et n'ont d'ailleurs pas attendu l'IA pour apparaître. Sur la manière de gérer les inégalités. Sur le rôle de l'école dans la préparation des talents de demain. Sur la fracture des intelligences qui risque de s'agrandir après celle de la fracture numérique.

Prenons du recul également sur ces peurs exacerbées sur la force des USA et de la Chine dans l'IA alors qu'elle est déjà un fait accompli dans les technologies clés du numérique d'aujourd'hui. Les comparaisons sont toujours abusives côté ordres de grandeur entre un pays de 68 millions d'habitants (la France), de 325 millions (USA) et de 1,4 milliards (la Chine).

Il faut juste se rappeller que nous sommes une puissance moyenne par rapport aux géants chinois et américains. L'Europe est un géant à leur hauteur mais fragmentée. L'approche gagnante ne peut être que transnationale, européenne et même latérale avec d'autres continents ou pays. Mais cette défragmentation est un casse-tête pour l'Europe.

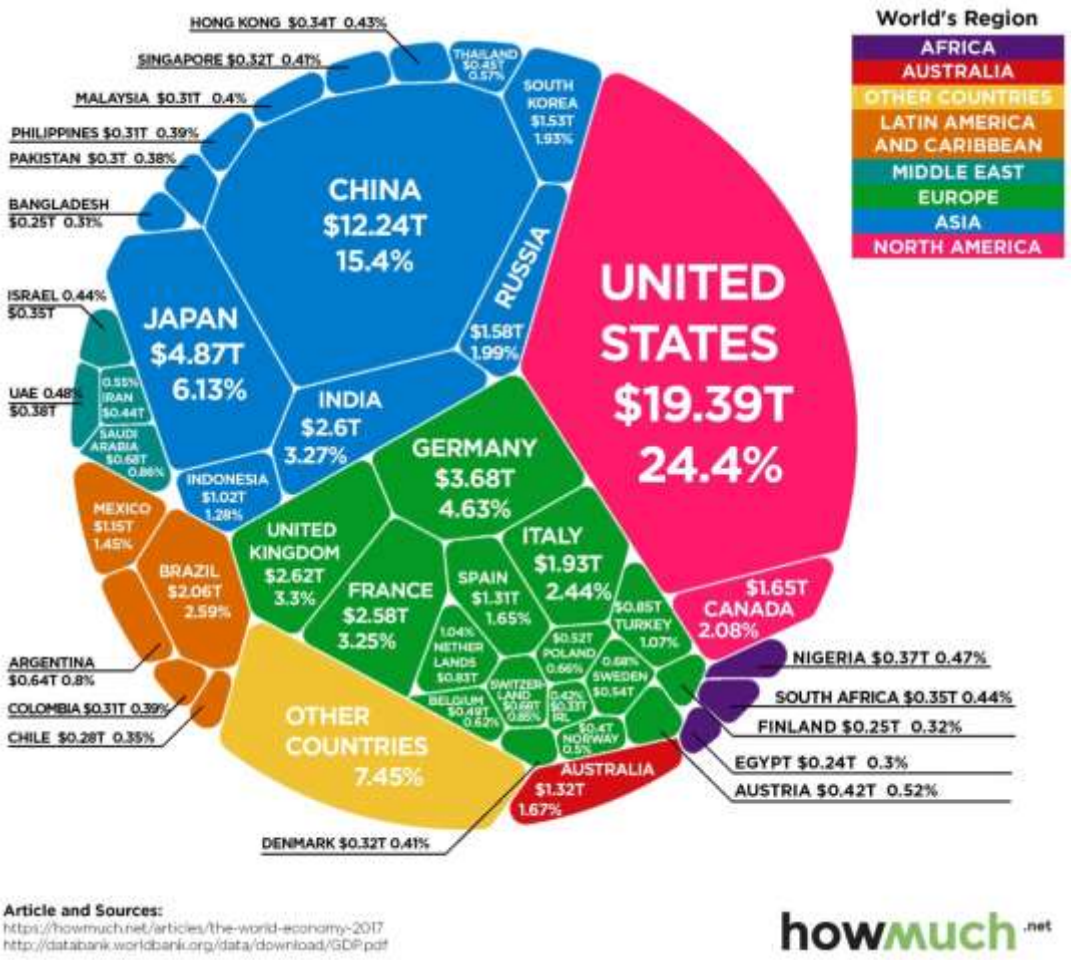
[1926] Voir Focus sur NaonedIA, le collectif nantais mettant en avant l'intelligence artificielle pour tous, de Johanna Diaz, juin 2018.

[1927] Auteur de l'ouvrage « Disruption » en 2019, qui décrit les bouleversements qui affectent les entreprises.

[1928] Avec cet ebook, je cible plutôt les entreprises, sans aucune prétention à toucher le grand public.

[1929] « L'intelligence artificielle est exposée au risque des technoprophètes », dans Le Monde, octobre 2018.

[1930] Avec une bonne moitié de prévisions qu'il fait siennes dans Laurent Alexandre : sept tendances qui annoncent l'intelligence artificielle de 2040, octobre 2018.

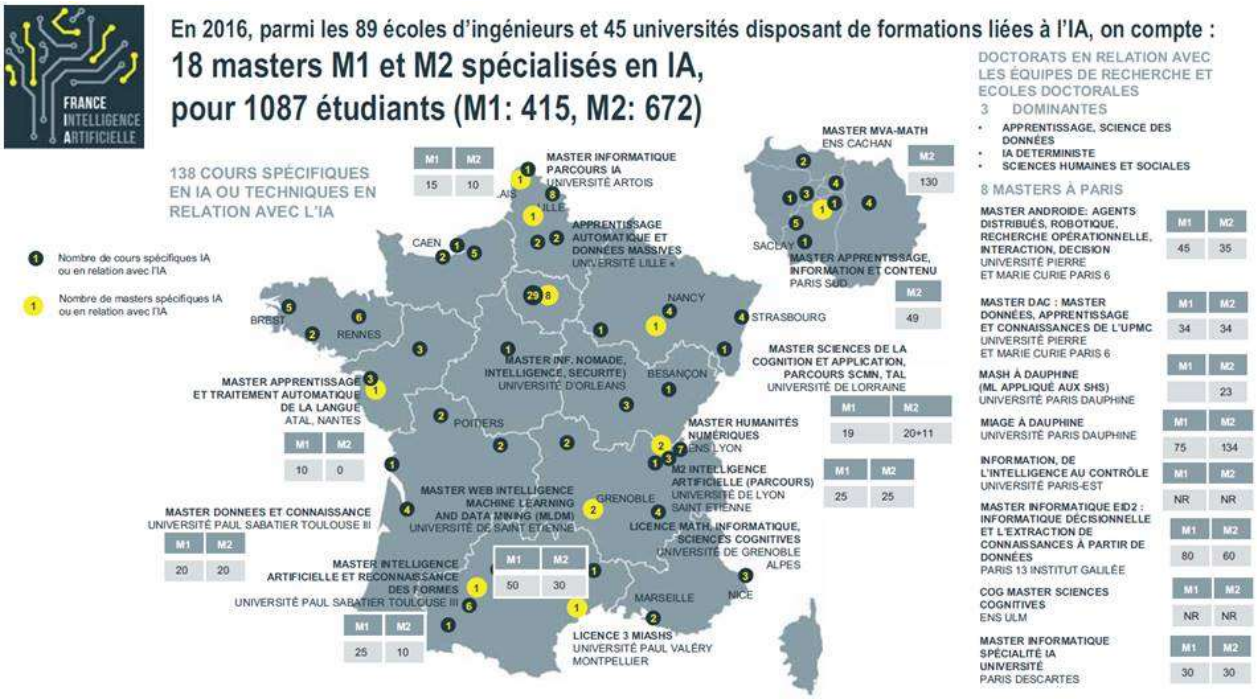


L'Histoire nous a construits fragmentés. Les réglementations européennes et l'Euro n'ont pas suffi à unifier les marchés.

Au-delà de questions sociétales et financières, c'est la principale raison de la quasi-absence de grands acteurs mondiaux du numérique en Europe (hormis SAP et Dassault Systèmes).

Éducation

Le Rapport France IA de 2017 faisait un inventaire des formations dans l'IA. Les Masters Spécialisés ne produisaient que 1087 étudiants par an, à comparer avec environ 30 000 développeurs formés par an et qui ne seraient déjà pas suffisants pour répondre à la demande. La situation est tendue sur ce marché. Les progrès de l'IA pourraient être ralentis dans certains pays par le manque de compétences. On ne paramètre par un réseau de neurones convolutifs ou récurrent en claquant des doigts.



D'où les propositions de ce plan et des plans suivants visant à augmenter le nombre d'étudiants formés à l'IA dans l'enseignement supérieur dans les divers plans gouvernementaux, qui insistent à juste titre sur le besoin de croiser ces formations avec des cursus non informatiques (santé, transports, etc)[1931].

Nombre d'initiatives de formations à l'IA sont venues du secteur privé[1932] sans compter l'autoformation avec des ressources en ligne, particulièrement appréciées par les plus jeunes[1933].

La question de l'éducation porte aussi sur toutes les autres filières d'enseignement professionnelles. Comment les rendre plus résilientes à la transformation ou à l'automatisation complète des métiers par l'IA ? C'est ce que traite Laurent Alexandre dans son ouvrage de 2017 « La guerre des intelligences » où il prône un enseignement non spécialisé favorisant la créativité.

Les formations professionnelles (CAP et BAC Pro) ou supérieures forment des spécialistes plus ou moins spécialisés. Ces spécialités n'ont pas empêché des tas de BAC+n de se réorienter dans divers chemins lors de leur vie professionnelle.

Les entreprises ont à la fois besoin de spécialistes prêts à l'emploi et de salariés qui s'adaptent rapidement au changement. C'est une attitude schizophrénique difficile à gérer. Il faut donc des spécialistes-généralistes... ! C'est-à-dire, des formations suffisamment généralistes et avec une spécialisation permettant quand même de démarrer quelque part. Nombre d'écoles d'ingénieurs et de commerce ont ainsi un tronc commun généraliste et une année de spécialisation. C'est un bon compromis qui mériterait d'être appliqué aux formations professionnelles.

[1931] L'un des exemples est l'annonce de la création d'un collège (BAC+2) au MIT pour former des jeunes à l'IA en la croisant avec d'autres disciplines et notamment la santé, les sciences politiques et la linguistique. Le collège sera financé à hauteur de $1B, probablement structuré sous la forme d'un *endowment* dont les intérêts feront tourner le collège sur plusieurs décennies. Il aura 50 enseignants. Le principal donateur est le CEO du fonds d'investissement Blackstone, Stephen Schwarzman, à hauteur de $350M. Voir Gift of $350 million establishes the MIT Stephen A. Schwarzman College, dans MIT News, octobre 2018 et M.I.T. Plans College for Artificial Intelligence, Backed by $1 Billion de Steve Lohr, octobre 2018.

[1932] Avec notamment le programme de formation proposé par Microsoft et Simplon.co à une cinquantaine de jeunes et de demandeurs d'emploi par an. Mais ce sont des formations BAC+2 assez légères. Elles rendent ces personnes employables dans les activités de services mais pas forcément au niveau attendu par les startups les plus exigeantes qui peuvent de leur côté avoir besoin d'au minimum des BAC+5 quand ce n'est pas des doctorants.

[1933] Comme les cours d'IA de Cornell ou de Stanford dont on peut facilement récupérer les supports en ligne avec par exemple Foundations of AI et Machine Learning.

La recherche de talents dans l'IA par les entreprises françaises est exacerbée par la fuite des talents vers les grandes entreprises étrangères qui proposent parfois des salaires mirobolants[1934]. Ceux-ci sont cependant souvent exagérés par l'extrapolation des cas les plus extrêmes[1935]. Des salaires compris entre $300K et $500K sont souvent évoqués, en oubliant qu'une part relève de stock-options. Ceci étant, les recrutements des entreprises de technologie créent un phénomène de vases communicant qui peut perturber l'activité des chercheurs, en ponctionnant leur capital humain trop brutalement. C'est vrai en France comme aux USA ou ailleurs[1936].

L'appel du marché a entraîné la création de nombreux cursus dédiée à l'IA dans l'enseignement supérieur[1937]. Citons par exemple le Master of Science en IA lancé par l'**EPITA** en septembre 2019, étalé sur 12 mois de formation complétés par 6 mois de stages en entreprise. Le tout complète une majeure déjà en place dédié à l'IA et aux sciences des données.

Inria lançait son Academy en septembre 2020, destinée à accompagner par la formation ses écosystèmes. Il s'agit surtout de former des ingénieurs et chercheurs à un ensemble de logiciels open source provenant d'Inria : Coq (un vérificateur de preuve formelle), Pharo (programmation orientée objet), SOFA (modélisateur pour la simulation physique, notamment dans le monde médical) et Scikit-learn (machine learning). Les formations sont délivrées en ligne avec les outils d'Inria Learning-Lab.

En octobre 2020, l'**ESTIA** (Ecole Supérieure Des Technologies Industrielles Avancées) lançait en partenariat avec l'Université de Sienne en Italie deux masters of science en IA et big data dénommés respectivement BIHAR et eBIHAR en présentiel à Biarritz et en enseignement à distance (MOOC). Il s'étale sur deux ans. Derrière se Master se cache la patte d'Oracle qui propose 6 cours en ligne sur les 12 qui sont proposés dans eBIHAR.

L'**Institut Mines Telecom Business School** a créé une offre de cours sur l'IA et les questions d'éthique. Cela devenait en 2020 une formation sur l'IA pour les profils non techniques : l'Executive Master « IA pour les managers innovants ». Il intègre la dimension sociétale et éthique dans les sujets traités et est articulé autour d'exemples issus des participants. Le tout devant permettre aux décideurs de discuter en connaissance de cause avec les équipes techniques auxquels ils ont à faire. Le tout sur 8 mois.

L'école de commerce **SKEMA** a créé un laboratoire en intelligence augmentée, le SKEMA AI Global Lab, qui aura son siège à Raleigh en Caroline du Nord et deux antennes, une sur Paris et une à Sophia-Antipolis. Il est dirigé par Thierry Warin. La **Faculté de Paris Descartes** ouvrait en septembre 2019 un cursus de formation pour un diplôme universitaire d'IA en santé. Notons aussi la création de l'**Institut DataIA** lancé en amont, en 2017, un Institut Convergence associé à l'Université Paris-Saclay.

Mais la liste continue avec des nombreuses nouvelles formations sur l'IA lancées en 2020 !

En septembre 2020, HEC Paris et l'Institut Polytechnique de Paris (IP Paris, qui rassemble l'Ecole Polytechnique, l'ENSTA Paris, l'ENSAE Paris, Télécom Paris et Télécom SudParis) lançaient **Hi! PARIS**, un centre interdisciplinaire de recherche et d'enseignement d'ambition mondiale consacré à l'IA et aux Sciences des données. Il allie éducation, recherche et innovation et s'appuie sur 300 chercheurs déjà en place dans les deux établissements. Le Centre est financé par des entreprises

---

[1934] C'est le cas pour ce centre de recherche Sequel à Lille en IA s'est vu dont ses talents ont été récupérés par les GAFA. Voir Au labo Sequel, des cerveaux aspirés par les GAFA par Laure Belot, septembre 2019.

[1935] Voir par exemple AI Geniuses Are Being Paid Over $1 Million At Elon Musk's OpenAI de Sam Shead, avril 2018.

[1936] Voir Tech firms try to address the risks the AI race poses for research, novembre 2018.

[1937] Voir l'étude Formations et compétences sur l'Intelligence Artificielle en France, lancée par Syntec Numérique et réalisée par l'OPIIEC (une instance paritaire associant syndicats professionnels comme Syntec et syndicats de salariés), 2019 (99 pages). Elle contient de très nombreuses sources de données sur l'emploi lié à l'IA en France et dans le monde. L'étude fait état de 7500 créations nettes d'emploi dans l'IA en France en 5 ans.

privées partenaires, les premières étant L'Oréal, Capgemini, TOTAL, Kering et Rexel, avec un budget annuel de 50M€. Il couvrira les secteurs de l'énergie, de l'environnement, de la défense, de la sécurité, de la santé, de la distribution, de l'industrie du luxe, des télécoms, de l'alimentation, de la finance et de l'assurance. Le Centre coridigé par Eric Moulines (IP Paris) et Thierry Foucault (HEC Paris) doit recruter une trentaine de nouveaux enseignants et 150 doctorants en IA.

Télécom Paris et l'**ENSTA** lançaient deux Mastères Spécialisés autour des données et de l'intelligence artificielle qui en couvrent à la fois les aspects techniques, socio-économiques et éthiques. Il est destiné à des étudiants de 30 à 40 ans et dure 8 à 10 mois, stage en entreprise compris.

L'**Aivancity School for Technology, Business & Society** est une nouvelle école privée spécialisée dans l'IA qui doit ouvrir en septembre 2021 à Cachan avec une première promotion de 300 étudiants. Elle est associée à l'ENSAM (Arts et métiers), à l'EPF ainsi qu'à Microsoft qui est décidément partout. L'école propose un programme de Master of Science en IA et données de type grande école qui recrute à BAC+2/3, un *bachelor of science* en IA appliquée qui recrute au niveau Bac puis un Master of Science en IA et business intelligence qui recrute au niveau Bac +4/5. L'école se positionne aussi à cheval sur les dimensions technologiques et sciences humaines.

Une nouvelle école de l'IA **Simplon.co / Microsoft** ouverte à Nancy, dédiée à la santé et devant accueillir 24 « apprenants » sur un parcours de 7 mois.

L'**Institut Lénoard de Vinci** de la Défense lançait un MBA Management de l'Intelligence Artificielle. C'est un parcours de 9 mois à mi-temps qui démarre à l'automne 2021.

Lancée en 2012, la **3W Academy** forme des développeurs web en ligne en seulement trois mois. Elle lançait en 2020 une formation de data scientist de trois mois de « téléprésentiel » en anglais en s'associant à **The AI Institute** (Paris) qui s'adresse à des diplômés d'écoles d'ingénieur ou ayant un Master universitaire scientifique. Elle couvre les mathématiques de l'IA, les réseaux de neurones, le deep learning, la vision artificielle ainsi que les outils de développement clés.

L'école d'ingénieurs **Télécom Physique Strasbourg** et le laboratoire de recherche **ICube** de l'Université de Strasbourg lancaient en 2020 la chaire industrielle Science des données et intelligence artificielle (SD&IA). Elle est financée par **Crédit Mutuel Alliance Fédérale et Euro-Information, Heppner, Hager Group, le groupe ÉS, Socomec** et **2CRSi**. Elle doit former des data-scientists et encourager la recherche.

Côté contenus US, voir l'**Open Source Data Science Masters**, un cursus de formation à un master en data science en open source en ligne[1938].

Recherche

Dans **The Entrepreneurial State** (2013) et dans **The Value of Everything : Making and Taking in the Global Economy** (2018), Mariana Mazzucato se bat contre l'idée selon laquelle le secteur privé prend des risques et l'Etat est conservateur et lent. Elle montre qu'au contraire, l'État – notamment américain – prend bien plus de risques et investit plus sur le long terme que toute entreprise privée.

C'est particulièrement vrai avec l'IA et encore plus vrai dans le cas de l'informatique quantique en France où pour l'instant, une majorité des risques sont pris dans la recherche publique, à l'exception rare d'Atos.

[1938] Voir http://datasciencemasters.org. Les ressources sont des vidéos en ligne, des exercices téléchargeables et une abondante bibliographie de textes disponibles en ligne. Voir aussi The Best AI Courses Leading to AI Certifications.

La recherche est à l'origine des grands progrès techniques dans l'IA, matérialisées ensuite par les offres des entreprises de toutes tailles. Geoff Hinton chez Google tout comme Yann Le Cun sont d'anciens chercheurs du secteur public au Canada et aux USA. Idem pour les fondateurs de DeepMind et de nombre de startups pointues dans le domaine.

De nombreux frameworks open source ont été créés par des chercheurs du secteur public comme **scikit-learn** qui a bénéficié et bénéficie encore de contributions d'Inria.

Au-delà de la recherche fondamentale, on a aussi besoin de plus de recherche appliquée dans tous les domaines et pas seulement dans le numérique « horizontal », ce qui explique par exemple l'implication d'Inria dans la santé.

La recherche française comprend quelques grands noms au CNRS, à Inria et ailleurs comme au CEA List : directeurs de recherche, patrons de laboratoires et autres, dont voici ci-contre un échantillon. On en trouve dans divers domaines : le machine learning, le deep learning, l'IA symbolique, les réseaux multi-agents, etc. Mais de nombreux chercheurs français ont été embauchés par les grands acteurs américains du numérique (en bleu dans la liste *ci-dessus*). Pas seulement par les GAFA, mais aussi par Nvidia, Microsoft et d'autres encore.

David Cournapeau, Scikit Learn, Enthought.
Jamal Atif, INS2I, LAMSADE, PRAIRIE – Paris.
Francis Bach, INRIA, DI-ENS – Paris.
Anne-Marie Kermarrec, Mediego.
Claude Berrou, IMT Atlantique.
Laurence Devillers, CNRS-LIMSI.
Françoise Soulié Fogelman, Tianjin University.
Yves Demazeau, AFIA.
Isabelle Bloch, LTCI Télécom ParisTech.
Raja Chatila, ISIR – Paris, éthique de l'IA.
Matthieu Cord, LIP6 – Paris.
Jean-Gabriel Ganascia, UMPC-LIP6.
Jean-Claude Heudin, Léonard de Vinci.
Chloé-Agathe Azencott, Ecole des Mines.
Michèle Sebag, CNRS-LRI.
Nicolas Ayache, INRIA, IA dans la santé.

Béatrice Daille, LS2N – Nantes.
Sébastien Konieczny, CNRS, CRIL – Lens.
Jérôme Lang, CNRS, LAMSADE – Paris.
Catherine Pelachaud, CNRS, ISIR – Paris.
Henri Prade, CNRS, IRIT – Toulouse.
Marie-Christine Rousset, LIG – Grenoble.
Marc Schoenauer, INRIA – Paris.
Isabelle Guyon, Université Paris-Saclay, machine learning.
Nozha Boujemaa, INRIA.
Thomas Schiex, INRA – Toulouse.
Nicolas Ayache, INRIA Sophia Antipolis, imagerie médicale.
Francis Bach, INRIA Paris, machine learning.
Cordelia Schmid, INRIA – Grenoble.
Jean-Philippe Vert, Mines ParisTech - ENS Paris.
Amélie Cordier, Hoomano, robotique.
Nicolas Vayatis, CMLA, ENS Paris-Saclay.

Yann LeCun, New York University, Facebook.
Antoine Bordes, Facebook Research Paris.
Camille Couprie, Facebook Research Paris.
Léon Bottou, Facebook Research, deep learning.
Jérôme Pesenti, IBM Watson, BenevolentAI, Facebook.
Emmanuel Mogenet, Google.
Rémi Munos, Google DeepMind, modélisation de la curiosité.
Vincent Vanhoucke, Google Research, IA et robotique.
Nicolas Papernot, Google Brain.
François Chollet, Google, Keras.
Olivier Bousquet, Google, machine learning.
Antoine Blondeau, Sentient.

Luc Julia, Samsung AI R&D.
Louise Naudin, Samsung AI à Paris.
Patrick Simard, Microsoft Research.
Nathalie Riche, Microsoft Research.
Yves Raymond, Netflix, machine learning.
Martial Hebert, Carnegie Mellon, robotique, vision.
Rodolphe Jenatton, Amazon, machine learning..
Nicolas Pinto, Apple, deep learning.
Timo Roman, Nvidia, deep learning.
Julie Bernauer, Nvidia, deep learning
Clément Farabet, Nvidia, hardware & infrastructure.
Erik Marcadé, SAP, head of AI.
Thierry Donneau-Golencer, Tempo AI, SalesForce.

Le plan proposé par la mission Villani de mars 2018 consistait à créer des 3IA (Instituts Interdisciplinaires d'IA), des instituts de recherche en IA pluridisciplinaires voisins du concept des IRT (Instituts de Recherche Technologiques) créés à partir de 2009 et à même de faire le pont entre la recherche fondamentale, la recherche appliquée et l'industrie. À l'issue d'un appel à manifestation d'intérêt puis un appel à projet, lancés par l'ANR, quatre 3IA ont été présélectionnés en 2018 : **MIAI**@Grenoble-Alpes (Grenoble, focalisé sur la santé, l'environnement et l'énergie, ce serait en 2021 le plus mature des 3IA), **3IA Côte d'Azur** (Nice, CNRS, Inria et l'Université Nice Côte d'Azur, focalisé sur la santé et le développement des territoires), **PRAIRIE** (Ile de France, sur la santé, les transports et l'environnement) et **ANITI** (Toulouse, Université Fédérale de Toulouse Midi-Pyrénées, sur les transports, l'environnement et la santé). Ces 3IA sont financés au maximum à hauteur de 100M€ par an via le PIA3 (Programmes d'Investissements d'Avenir).

Dans un document d'origine chinoise[1939], j'ai trouvé cet étonnant classement rarement utilisé qui place le CNRS et Inria en seconde et troisième position mondiale dans les centres de recherche publique en termes de publications. Par contre, ils disparaissent du palmarès en termes de brevets, laissant cette place au CEA.

De nombreuses entreprises étrangères ont un laboratoire de recherche en IA basé en France. Nous avons celui d'IBM du plateau de Saclay, déjà cité. Puis le Coréen **Naver** à Grenoble, qui a repris en 2017 les activités de recherche de Xerox sur place. **HP Enterprise** a aussi un centre compétence en IA à Grenoble qui travaille notamment avec le CEA-Leti dans le cadre du 3IA MIAI, et d'une chaire travaillant sur les questions environnementales comme la modélisation de l'effet du changement climatique sur la faune. **Facebook** a son laboratoire FAIR à Paris depuis 2015. Et **Microsoft** et Inria partagent depuis 2005 un laboratoire de recherche conjoint à Orsay.

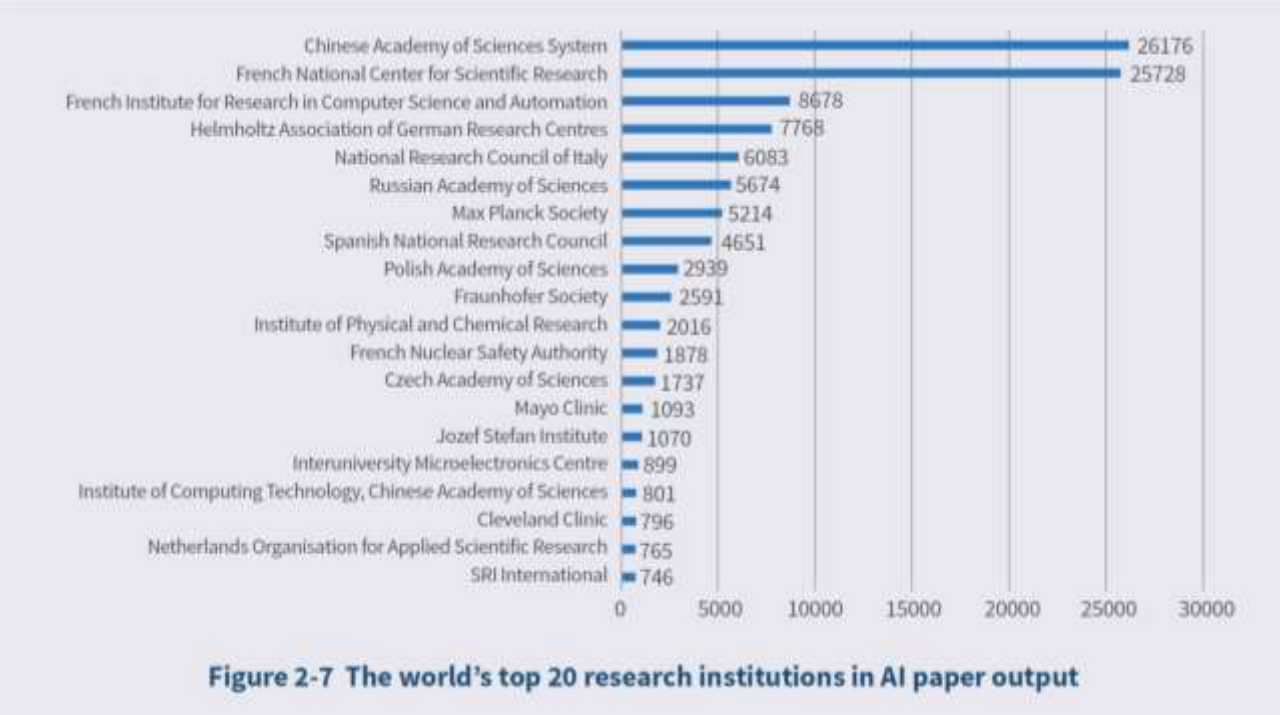


Figure 2-7 The world's top 20 research institutions in AI paper output

Économie et social

Les gouvernements font face à une *perfect storm* en perspective : une révolution technologique en accéléré qui pourrait rapidement bouleverser l'équilibre bancal actuel du marché de l'emploi. Même avec les prévisions les plus optimistes qui anticipent un déficit net de 6% d'emplois d'ici une dizaine à une quinzaine d'années, cela donnerait du grain à moudre.

Supporter une augmentation moyenne de 50% du chômage ne serait pas facile à absorber, même si les prévisions sont très élastiques au niveau de leur échéance et si des pays comme la Grèce et l'Espagne sont déjà passés par là pour d'autres raisons après leurs crises de la dette. Les syndicats commencent à s'en inquiéter, observant l'impact des gains de productivité liés à l'IA sur l'emploi ou sur les objectifs données aux salariés.[1940]

[1939] Voir China AI Development Report 2018, juillet 2018 (122 pages).

[1940] La CGT organisait un colloque sur l'IA en novembre 2018. J'y intervenais sur le sujet des stratégies industrielles de l'IA, à la suite de Bertrand Pailhes, coordinateur de la stratégie IA de l'Etat, qui présentait la feuille de route du gouvernement. Voir L'intelligence artificielle est l'affaire de toutes et tous de Marie-José Kotlicki, secrétaire de la Ugict CGT, la branche de la CGT pour les ingénieurs, cadres et techniciens. Elle fait un compte-rendu des débats de ce colloque. On peut compléter cela avec le compte-rendu de Christophe Alix dans Libération dans Intelligence artificielle : pour la CGT, le binaire de la guerre suite à mon intervention dans le Congrès de la CGT sur l'intelligence artificielle le 6 novembre 2018.

Quelles politiques économiques adopter ? Elles tournent toujours autour du développement économique avec les moyens à disposition des États pour accompagner les entreprises, en général par le financement de l'amont de l'innovation et de la recherche, et aussi en développant le tissus économique des startups.

L'un des points clés est d'être exportateur de technologies plutôt que simple consommateur. Si la valeur ajoutée de l'IA et les robots viennent d'un nombre réduit de pays, les autres seront toujours désavantagés comme nous le sommes déjà aujourd'hui dans de nombreux pans du numérique, surtout grand public.

Dans **Economic Report or The President**[1941], le rapport annuel 2016 sur l'économie de la Maison Blanche publié en janvier 2017 à la fin de l'administration Obama, on découvrait qu'aux USA et en 2013, les startups avaient créé 2 millions d'emplois et les entreprises traditionnelles 8 millions. Donc 20% !

Une proportion énorme sachant que dans le même temps, l'économie française a plutôt détruit des emplois et les startups n'en ont probablement créé que quelques dizaines de milliers tout au plus. Et surtout : la moitié de la R&D fédérale US est dédiée à la défense ! Au milieu des années Reagan (1985), elle en représentait les deux tiers !

Cela explique pourquoi tant de projets autour de l'IA sont financés par la l'agence de l'innovation du Pentagone, la DARPA[1942]. Y compris trois défis lancés en 2004, 2005 et 2007 sur la conduite automatique, qui ont dynamisé les équipes de recherche de nombreuses universités sur le sujet. Nombre de ces équipes ont été ensuite recrutées par Google pour ses différents projets de voitures automatiques.

En avril 2017, le Secrétaire du Trésor de Donald Trump, Steve Mnuchin, ancien de Goldman Sachs, affichait un optimisme étonnant[1943], affirmant dans un débat qu'il ne voyait pas de menace sur l'emploi causée par l'IA avant 50 ou 100 ans. Cela lui a valu le sobriquet d'*AI denier*, comme un *climate change denier*. On est passé d'un souci mesuré à une insouciance coupable, mais à la hauteur des compétences de l'actuelle administration américaine.

Et cela n'a pas changé depuis un an et demi. Même si on peut être optimiste, il est probable que Steve Mnuchin n'avait pas vraiment réfléchi à la question !

En France, le débat politique autour de l'IA a connu un tournant « social » pendant la présidentielle 2017. Il a contribué à la mise en avant de propositions de revenu minimum ou de base[1944]. Benoît Hamon et Jean-Luc Mélenchon le justifiaient avec les risques de robotisation des métiers. Une charrue mise avant les bœufs, même si les politiques sont parfois des bœufs et sont bien devant.

L'élection de Donald Trump par les fameux blancs de la middle class de la *rust belt* devait aussi à la désindustrialisation ce ces États et à un décalage mal vécu entre ces États qui s'appauvrissent et la Silicon Valley qui s'enrichit sans discontinuer.

[1941] Voir Economic Report to the President, January 2017 (599 pages).

[1942] En France, cette mission était historiquement dévolue à la DGA. Mais le Ministère des Armées a créé en septembre 2018 sa propre agence de l'innovation de la défense, dirigée par Emmanuel Chiva et visiblement rattachée à la DGA. Voir Création de l'Agence de l'innovation de défense et nomination d'Emmanuel Chiva au poste de directeur, septembre 2018.

[1943] La transcription de son intervention est sur Read the full transcript from Treasury Secretary Steve Mnuchin's interview, mai 2017.

[1944] Voir Today's Artificial Intelligence Does Not Justify Basic Income de Vincent Conitzer, octobre 2016, et c'est toujours valable deux ans plus tard et Les secrets bien gardés du revenu universel de Diana Filipova, août2016, qui met en avant la cacophonie idéologique autour du revenu universel.

Se pose aussi la question de la **politique fiscale**, notamment vis-à-vis des GAFAMI qui sont déjà accusés d'évasion fiscale, avec leur statut d'agent commercial appliqué à leurs filiales. L'IA pourrait accentuer le phénomène de migration de valeur sur les plateformes que ces GAFAMI contrôlent. Cela relance aussi les procédures antitrust en cours, pilotées par l'Union Européenne, sans compter les mouvements politiques allant dans ce sens aux USA.

Autre sujet de débat, celui de la **taxation des robots**, proposée notamment par Bill Gates[1945]. Pourtant, la taxation des robots est un système bien compliqué, inadapté et n'a pas de sens sans une fiscalité internationale hétérogène. Ce n'est pas plus malin de que de taxer des machines à tisser ou les tableurs Excel !

Si on taxait les robots, il faudrait alors taxer tous les outils matériels et immatériels qui ont amélioré la productivité du travail depuis quatre millénaires : les tracteurs qui ont permis le développement de l'agriculture intensive et de passer, pour prendre la France en exemple, d'une population agraire de 36% des salariés en 1946 à moins de 2% après les années 2000, les logiciels qui ont permis de se passer de secrétaires dans nombre d'entreprises, les machines-outils dans les usines, les tableurs qui ont réduit les besoins en comptables, les moteurs de recherche et l'information en ligne qui ont réduit l'attrait des bibliothèques et plein d'autres évolutions du même genre.

Et puis, pourquoi donc taxer les robots physiques alors que l'IA immatérielle pourrait supprimer encore plus d'emplois que les robots logiciels, tout du moins dans les pays développés ? Et si on taxait les robots, cela réduirait l'intérêt économique de rapatrier des usines dans les pays développés, le contraire d'une réindustrialisation. Faudrait-il faire la distinction entre les robots d'usines et les robots humanoïdes ?

La question ne se pose pas pour l'instant. Plutôt qu'inventer une taxe spécifique pour les robots, les États pourraient commencer par appliquer sérieusement les taxes génériques qui concernent les entreprises.

Si un État met en place une taxation des robots immatériels qui détruisent des emplois, il y en aura toujours d'autres pour accueillir les entreprises concernées et leur servir de paradis fiscal. Il est donc plus important d'homogénéiser la fiscalité que d'en inventer une nouvelle. Si des robots suppriment massivement des emplois, cela améliorera la profitabilité des entreprises et il suffit alors de les taxer correctement sur leurs bénéfices plutôt que sur leur outil de travail qui est une structure de coût et pas de profit.

Une taxe sur les robots appliquée uniquement en France ne ferait que pénaliser l'industrie française par rapport aux autres pays qui font appel à la robotisation, y compris en Asie.

Sans compter le fait qu'aujourd'hui, diverses études montrent que les pays les plus robotisés sont ceux qui se développent le mieux !

En taxant simplement les profits, comme on le fait aujourd'hui, on taxe l'ensemble des sources d'économies d'échelle et pas seulement la robotisation. Les entreprises qui sont et seront les plus robotisées auront les meilleurs profits, c'est tout. Il suffit de taxer l'eau qui coule à la fin du circuit économique que dans les multiples robinets qui font tourner l'entreprise.

Il vaut mieux investir dans la formation et les compétences des gens pour les aider à créer des robots, à les installer, les maintenir, les piloter, les superviser, à gérer des projets les intégrant. On ne résiste pas à l'innovation. On s'y adapte et on aide les gens à s'y adapter.

Qui plus est, le risque de pertes d'emplois lié à l'automatisation est plus fort dans les métiers non physiques que dans les métiers physiques. Un expert-comptable est plus menacé par l'IA qu'un kinésithérapeute ou une sage-femme ne le sont par des robots.

[1945] Voir Here's how Bill Gates' plan to tax robots could actually happen de Malcolm James, mars 2018, Bill Gates Is Wrong That Robots And Automation Are Killing Jobs de James Bessen, février 2017 et What's Wrong With Bill Gates' Robot Tax de Noah Smith, février 2017.

Dans le cas de la robotisation dématérialisée, que faudrait-il taxer ? Les logiciels et le cloud ? Ils sont de moins en moins chers. Bref, on tourne en rond.

Enfin, dernier point : l'impact des outils de traduction automatique sur la construction européenne. L'un des écueils de l'Union Européenne est sa grande diversité linguistique qui complique la communication. Est-ce que la traduction automatique va permettre de passer outre ces barrières ? C'est possible. Mais la fragmentation du marché européen n'est pas que linguistique. Elle est aussi forte dans presque tous les secteurs d'activité. Les effets de levier économique varient d'un pays à l'autre : les médias, les banques, les télécoms, les *retailers* et les services en général sont le plus souvent locaux. Les USA sont beaucoup plus homogènes de ce point de vue-là et ils sont les principaux pourvoyeurs de produits et services qui sont communs à toute l'Europe, de Google à Amazon.

Souveraineté

Les usages de l'IA posent évidemment des questions clés de ce côté-là.

Sans l'IA, de nombreux services Internet jouent déjà le rôle de régulateurs privés de l'Internet, qu'il s'agisse de Google Search ou de l'algorithme d'alimentation de votre timeline sur Facebook. Avec l'IA, la situation va se corser car les résultats de ces outils vont de moins en moins dépendre d'algorithmes qui peuvent être décortiqués par rétro-ingénierie et de plus en plus de solutions à base de deep learning qui ne sont pas facilement auditables.

Les pouvoirs publics en sont à réclamer des solutions techniques permettant d'expliquer les algorithmes. Pourtant, ceux-ci sont compréhensibles par les spécialistes ! C'est le caractère automatisé de la création des « feature maps » intermédiaires des réseaux de neurones convolutifs qui déroute.

Second enjeu, l'émergence de nouvelles fragilités en termes de sécurité avec la capacité de tromper les techniques de deep learning, notamment dans la reconnaissance d'images ou via les données issues des objets connectées. Cela entraîne un besoin de sécurisation encore plus poussée des infrastructures stratégiques, de défense et de cybersécurité des Etats.

Les États devront se doter de leurs propres solutions d'IA spécifiques pour préparer leurs décisions stratégiques et anticiper celle des autres Etats. Cela relève encore de la bordure de la science-fiction mais beaucoup moins que les thèses de la singularité.

Les outils du renseignement et de la société de surveillance exploiteront de plus en plus l'IA, notamment pour faire des recoupements d'information pour identifier des profils suspects d'Internautes.

L'ancien Directeur de la NSA et de la CIA pendant la présidence Bush 43, Michael Hayden, évoquait en 2016 la séparation juridique sur la vie privée et de la sécurité en Europe[1946].

Elle est gérée de manière globale aux USA, tandis que dans l'Union Européenne, la vie privée est réglementée au niveau communautaire et la sécurité au niveau des pays. Ce qui créé un handicap pour les Etats.

La souveraineté des États sera aussi remise en cause en cas de transformation très radicale de certains métiers. L'impact pourrait être grand dans certains pays où des activités sont délocalisées, comme en Inde. Ces pays seront probablement affectés par la robotisation de ces activités, comme celles qui seront affectées par la RPA (Robotic Process Automation). Dans le même temps, ils pourraient bénéficier des technologies de traduction automatique pour couvrir plus de marchés !

Régulation

L'IA soulève de nombreuses questions juridiques qui font l'objet de débats depuis plusieurs années et notamment au niveau européen.

[1946] Il me semble que c'était dans l'intervention Inside the NSA: An Evening with General Michael Hayden, datant de 2014.

En France, le secrétaire d'Etat en charge du numérique, Mounir Mahjoubi, avait lancé en juillet 2018 les États généraux des nouvelles régulations numériques visant à définir une position française dans les négociations européennes sur les sujets de la régulation numérique et en particulier celles qui sont liées à l'IA. C'est un moyen de formaliser la collecte des retours des entreprises et des start-ups.

Les retours connus portent surtout sur les critiques concernant les abus de position dominante des GAFA et les blocages opérationnels que peuvent générer l'application du RGPD. Du côté de l'IA, les startups sont surtout en demande de données pour alimenter leurs solutions et en particulier dans le domaine de la santé, et cela ne relève pas forcément des compétences européennes.

Les GAFAMI prennent des initiatives ambiguës qui visent en général à limiter les tentations de régulation des États. Ils affichent une « responsabilité », parfois de façade comme chez Facebook, qui prétend relever d'une autorégulation[1947].

La **personnalité juridique des robots** est un autre sujet discuté à l'échelle européenne. L'avocat Alain Bensoussan suggère, avec enthousiasme[1948], de créer un véritable droit des robots, situé entre le droit des biens et des personnes[1949]. S'y ajouterait la création de référentiels robotiques aux niveaux éthiques, culturels et normatifs. Ce droit comprendrait les règles générales applicables à tous les types de robots, les règles applicables à des robots spécifiques comme les véhicules autonomes, les robots chirurgiens autonomes ou les robots de services humanoïdes.

Un robot aurait une identité constituée d'un numéro[1950] et même une assurance. Avec l'ambiguïté liée au fait que le robot et le logiciel qui l'animent ne sont pas étroitement associés comme dans les êtres humains ou les animaux, le dernier pouvant tourner dans le cloud et servir plusieurs robots à la fois, pouvant aussi être hacké et mis à jour.

Qui plus est, la responsabilité d'un robot en cas d'accident associe son concepteur, le logiciel, les données qui l'alimentent et l'influencent, dont les actions des humains qui l'entourent et les données environnementales. Les responsabilités ne sont plus individuelles, mais des chaînes complexes de responsabilités.

200 experts européens ont émis de sérieuses réserves sur la question en avril 2018 via une lettre ouverte[1951]. La notion de personnalité juridique du robot a été abandonnée par la commission européenne. Il subsistera trois types de responsabilité : celles des concepteurs, celle des sociétés et personnes qui exploitent les robots, et en dernier lieu, celle de ceux qui interagissent avec.

Le fonctionnement des robots, et en particulier des véhicules autonomes, devra être auditable pour déterminer les responsabilités en cas d'accidents.

[1947] Voir Les géants de l'intelligence artificielle sauront s'autoréguler, dit Google, AFP, 2018. Et plus récemment, The invention of ethical AI - how Big Tech Manipulates Academia to Avoid Regulation par Rodrigo Ochigame, décembre 2019 puis Comment les géants de la tech manipulent la recherche sur l'éthique des IA par Antoine Hasday, janvier 2020.

[1948] Voir son intervention à TEDx Paris en octobre 2015 : De l'urgence d'un droit des robots. Alain Bensoussan a même créé une Association pour le Droit des Robots en 2014.

[1949] Alain Bensoussan a aussi lancé divers servcies juridiques en ligne basés sur de l'IA avec sa propre équipe de développeurs. Voir https://www.alain-bensoussan.com/avocat-intelligence-artificielle/. Il propose notamment une solution de justice prédictive.

[1950] Mais peut-être aussi accompagné de la version des logiciels qui l'animent, de ses capteurs, de leur état, et des données qui alimentent ses logiciels et peuvent affecter son comportement ! Le robot ne sera pas Skynet mais sa connexion à de nombreux services créera un système fortement maillé difficile à isoler.

[1951] Voir Une lettre ouverte pour refuser la « responsabilité juridique » des robots, de Rémy Demichelis, dans Les Echos, avril 2018. Ainsi que « Accorder des droits à une machine, c'est une pente dangereuse » sur la position de Nathalie Nevejans qui est opposée à cette idée de personnalité juridique accordée aux robots, octobre 2018.

Les premiers accidents impliquant des véhicules autonomes ou semi-autonomes ont ainsi permis à un bon niveau de déterminer les fautes entre concepteurs (véhicules n'ayant pas tenu compte de la détection de personnes sur la route comme dans l'accident en Arizona avec un véhicule Uber semi-autonome), conducteurs (n'ayant pas respecté les règles de vigilance) et personnes ayant pu provoquer ces accidents à partir d'autres véhicules (non-respect de priorité…).

La réglementation de l'usage des véhicules autonomes se posera, dont la question de la fameuse gestion des dilemmes lorsqu'un véhicule autonome doit choisir entre deux formes d'accidents et de dommages corporels. Comme en cas de choc inévitable sur une voiture parmi deux, faut-il choisir de percuter la voiture avec un seul passager ou celle avec plusieurs passagers ? Est-ce que la réglementation devra s'appliquer à ce genre de choix moral ? Probablement.

Même si ces choix seront très rares pour ces véhicules. On aura besoin de connaître l'avance le comportement des véhicules pour savoir comment interagir avec. Certains passants s'amusent aujourd'hui à tester la détection de piétons des véhicules autonomes comme les Navya et ont tendance à les empêcher de fonctionner. Un test qu'ils sont moins tentés de faire avec un tram lancé à toute vitesse, sachant qu'il n'aura pas le temps de s'arrêter.

Le second point clé concerne la **protection de la vie privée**, qui risque d'être encore plus mise à mal avec l'IA qui va accumuler et croiser de nombreuses données très personnelles. Avec l'application du RGPD, la réglementation européenne en vigueur depuis mai 2018, les entreprises européennes doivent se conformer à des règles plus strictes sur la protection des données privées. Cela pourrait gêner le déploiement de solutions d'IA grand public et favoriser les GAFA, même si ces derniers devront respecter la même réglementation en Europe. Mais des voix se lèvent aux USA pour réclamer la création d'un RGPD pour les USA[1952], y compris celle de Tim Cook, le CEO d'Apple.

Le droit à l'oubli qui est inscrit dans la loi « République numérique » d'octobre 2016 (dite loi « Lemaire ») devra donc s'appliquer aussi aux IA et aux robots de services à qui ont devrait pouvoir demander de ne pas se souvenir d'événements.

On peut se demander comment pourrait fonctionner le droit à l'oubli dans un réseau de neurones complexe dont les paramètres ont été affectés par le comportement d'un utilisateur donné. Faudrait-il réentraîner tout le réseau à partir de zéro pour éviter que celui-ci reconnaisse un utilisateur en fonction de son comportement ?

Quid sinon de l'application du Premier Amendement qui régit la liberté d'expression aux USA, à des robots logiciels ?

Le droit « case law » des USA est très différent du droit romain qui sévit en Europe. Aux USA, une bonne partie du droit provient de la jurisprudence. Il préempte peu l'innovation. En Europe et en France, le droit romain domine et cherche parfois à précéder l'innovation.

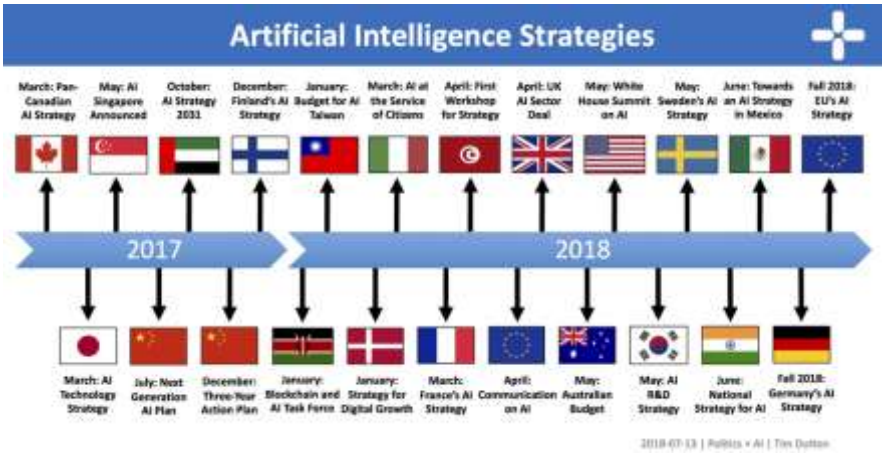
Cette différence d'approche a un impact sur la réglementation applicable aux innovations technologiques. Elle favorise plutôt les Américains ! Quant aux Chinois, l'ordre collectif prime sur les libertés individuelles et le droit est donc inversé. La personne se soumet donc tant bien que mal à une société de la surveillance, et notamment de la vidéosurveillance et de la notation comportementale.

Géopolitique de l'IA

L'IA est un véritable enjeu industriel et politique à l'échelle planétaire. Elle amplifie le jeu de la concurrence entre les trois grands continents technologiques : l'Amérique du Nord, l'Asie et l'Europe.

[1952] Voir Privacy group calls on US government to adopt universal AI guidelines to protect safety, security and civil liberties, octobre 2018. L'initiative venait de l'Electronic Privacy Information Center (EPIC).

Depuis 2015, tous ces pays, essentiellement ceux de l'OCDE, rivalisent d'annonces et d'investissements pour faire d'eux des « champions » de l'IA[1953]. La Chine l'a également fait, marquée notamment par le Spoutnik moment de la victoire d'AlphaGo de DeepMind face aux meilleurs joueurs mondiaux du Go, tous asiatiques.



Les pays utilisent un panachage des mêmes artifices : des investissements publics dans la formation et la recherche, la stimulation de l'entrepreneuriat, son financement, et enfin, la modernisation de l'État aussi bien dans les services publics que dans les activités régaliennes comme celles du renseignement, de la sécurité et de la défense[1954].

Les pays se pressent pour devenir leaders de l'IA tout en craignant les effets délétères de l'automatisation sur l'emploi. Dans les pays occidentaux, la question de l'éthique de l'IA est aussi mise en avant face aux Américains qui ont tendance à rendre le choix binaire entre céder aux Chinois ou mettre cette éthique au placard.

Les ambitions se situent au niveau des zones économiques comme l'Europe, sur les pays, sur les clusters technologiques et sur les capitales. Les pays « tier 2 » comme les grands pays européens se comparent souvent avec les USA et la Chine. Quantitativement, cela ne peut que leur être défavorable.

Nombre de pays annoncent des montants d'investissements publics dans l'IA. Après en avoir fait le tour, je recommande toujours de les prendre avec des pincettes transparentes. En effet, ces montants sont toujours des compléments à des investissements existants qui ne sont presque jamais inventoriés. Parfois, ils remplacent des montants existants. Bref, on ne dispose pas toujours d'une vision cohérente de l'ensemble des investissements publics et privés dans l'IA.

De plus, les montants sont presque toujours pluriannuels. Il faut les ramener à des montants annuels pour pouvoir les comparer d'un pays à l'autre, puis, les comparer au prorata par rapport au PIB des pays[1955]. En se rappelant par exemple que le PIB des USA équivaut à 7,8 fois celui de la France ! Comme pour l'entrepreneuriat et les startups, on voit fleurir des classements par pays sur l'IA. Ils sont souvent indexés sur les publications scientifiques et/ou sur les startups d'IA créées[1956].

Le politique est bien en mal pour anticiper le futur. Il s'est emparé du sujet mais n'est pas mieux armé que les prospectivistes et analystes qui sont tout autant désemparés.

[1953] Voir par exemple L'intelligence artificielle, une course mondiale à l'innovation (non daté).

[1954] Voir l'inventaire de An overview of national AI strategies, juin 2018 qui couvre 31 pays et leurs initiatives autour de l'IA. Je m'en suis inspiré pour la rédaction de cette partie.

[1955] Par exemple, comparer un investissement public annuel d'un pays avec le milliard de dollars qui vient d'être mis de côté pour financer un collège au MIT n'a pas de sens. Ce $1B d'origine privée va alimenter un fonds dont les intérêts annuels serviront à faire tourner ledit collège. Cela donnera donc aux alentours de $30M par an. Si un pays investi un montant voisin chaque année dans un laboratoire ou un établissement d'enseignement public, il sera donc équivalent à ce milliard de dollars !

[1956] Voir 5 Countries Leading the Way in AI de Bruno Jacobson, janvier 2018. Il liste cinq pays : Chine, USA, Japon, UK et Allemagne. Ils sont évalués selon leurs publications scientifiques dans l'IA. Sur 20 ans et en 2017, la France est septième selon le Scimago Journal & Country Rank. Voir le classement dans l'IA sur 20 ans http://www.scimagojr.com/countryrank.php?category=1702 et le même sur 2017 : https://www.scimagojr.com/countryrank.php?category=1702&year=2017.

Surtout lorsque l'on essaye d'intégrer le temps long dans la vision, comme sur l'impact des technologies sur le fonctionnement des démocraties, sans compter les inévitables crises financières cycliques qui secouent le Monde[1957].

En attendant, les startups de l'IA fleurissent dans presque tous les pays du monde qui veulent leur part d'un gâteau incertain[1958]. L'IA est devenue une sorte de commodité comme les logiciels.

USA

Les USA sont nettement le leader mondial de l'IA, en tout cas, dans le mix recherche, industrie et startups. L'écosystème du pays bénéficie notamment de la prodigalité des budgets de la défense et du renseignement qui alimentent nombre de projets de recherche appliquée dans le privé et les universités. Le capital risque est aussi très dynamique pour financer les startups de l'IA. En tout $130B en 2018 !

La dominance économique des GAFAMI et leur capacité d'acquisitions de startups leur apporte la puissance industrielle. Celle-ci est aussi soutenue par le rôle que jouent des acteurs tels qu'Intel et Nvidia. Malgré cette supériorité économique des GAFAMI, certains Américains s'inquiètent sérieusement de la montée en puissance de la Chine dans l'échiquier mondial.

La prise en main de l'IA par l'exécutif américain a été progressive. La Maison Blanche avait publié quelques rapports et plans sur l'IA du temps de la présidence de Barack Obama en 2016[1959].

Le plan d'octobre 2016 comprenait le financement de la recherche long terme[1960], le développement de méthodes de relations entre humains et IA, le traitement des questions d'éthiques, légales et sociétales de l'IA, assurer la sécurité des solutions d'IA, la création de jeux de données publics pour l'IA, celle d'outils d'évaluation et de benchmarks de solutions d'IA et enfin, l'évaluation les besoins en formation. Somme toute des questions assez classiques. Aucune mesure ne concernait l'entrepreneuriat et les startups car le système de financement privé US fonctionne bien sans intervention publique directe[1961].

[1957] Voir Technological Revolutions Bring About Fascism. Who Will Save Us This Time? de Nicolas Colin dans Forbes, octobre 2018. Il y décrit le lien entre grandes transitions technologiques et émergence du fascisme en faisant un parallèle avec les années 1930.

[1958] Voir Top-10 Artificial Intelligence Startups in Spain, février 2019, Top-10 Artificial Intelligence Startups in Switzerland, janvier 2019, Top-10 Artificial Intelligence Startups in South America, janvier 2019, Top-11 Artificial Intelligence Startups in Indonesia, janvier 2019, Top-10 AI Startups in Central and Eastern Europe, juillet 2019 et Top-10 Artificial Intelligence Startups in Africa, avril 2019.

[1959] Voir The Administration's Report on the Future of Artificial Intelligence, octobre 2016, National Artificial Intelligence Research and Development Strategic Plan, octobre 2016 (48 pages) et Artificial Intelligence, Automation, and the Economy, décembre 2016.

[1960] Les USA ont investi $1,1B de crédits fédéraux en 2015 dans la recherche civile en IA. Le Pentagone aurait dépensé $7,4B dans l'IA et ses technologies de support comme comme le big data et le cloud pour ce qui est des crédits non classifiés. Ce qui dépend de l'appareil du renseignement (NSA, CIA, DIA, NRO) n'est pas documenté publiquement.

[1961] Ce plan passé quasiment inaperçu dans les médias US avait été publié dans la dernière ligne droite de la campagne présidentielle américaine, entre les révélations de l'Access Hollywood Tape de Donald Trump et les déclarations de James Comey relançant pour 10 jours l'enquête sur les mails d'Hillary Clinton.

A partir de janvier 2017, l'administration Trump n'a pas bougé le petit doigt sur le sujet de l'IA pendant un an et demi. Donald Trump a passé plus de temps à relancer - en vain - la production de charbon que sur l'ensemble des technologies et sciences[1962]. L'écosystème américain a alors commencé à s'en inquiéter, surtout au vu des avancées de la Chine dans le domaine. Cela devenait un enjeu stratégique et de souveraineté[1963]. A la décharge de la Maison Blanche, tant celle de Barack Obama que de Donald Trump, celle-ci se contente souvent de généralités dans la création de tels plans.

Les détails financiers sont laissés à l'initiative du Congrès et la mise en œuvre des plans aux grandes agences fédérales qui distribuent ensuite une bonne part de leurs financements au secteur privé, dont la recherche des universités.

Un sommet de l'IA était organisé à la Maison Blanche en mai 2018. Le conseiller scientifique adjoint de l'époque, Michael Kratsios, y présentait l'approche dans l'IA de l'administration Trump avec un plan en quatre points comprenant, en particulier, la volonté de supprimer les barrières à l'innovation en assouplissant les règles sur l'expérimentation de véhicules autonomes et de nouvelles méthodes de diagnostic dans la santé[1964].

L'un des moteurs fédéraux de l'innovation dans l'IA aux USA est la **DARPA** qui a lancé divers défis technologiques comme ceux des véhicules autonomes entre 2005 et 2007 et puis, plus récemment, autour de l'IA explicable (**XAI**)[1965].

Il faut aussi prendre en compte l'action parallèle d'une autre agence fédérale moins connue, l'**IARPA** qui est la DARPA du renseignement, rattachée au Directeur National du Renseignement. Elle cofinance également des projets de recherche fondamentale et appliquée dans les Universités, startups et entreprises, en particulier par le biais de Challenges.

Elle est particulièrement intéressée par les applications de l'IA dans le traitement du langage. Par exemple, le **Mercury Challenge** est un défi pour prédire le futur[1966] !

Donald Trump signait un décret le 11 février 2019 pour créer l'**American AI Initiative**. Cette initiative était une déclaration d'intentions qui proposait des réallocations budgétaires dans les agences fédérales[1967], le Congrès devant prendre le relais pour allouer des ressources et budgets.

[1962] Les relations entre le Président Trump et les sciences ont toujours été quelque peu distendues. Voir Is Donald Trump anti-science? The data says yes, par Mark Quigley et al, janvier 2020.

[1963] Voir France, China and the EU all have an AI strategy. Shouldn't the US?, de John Delaney, mai 2018, The Trump administration plays catch-up on artificiel intelligence, de Tom Simonite, mai 2018, Trump administration takes a hands-off approach to AI, de Kris Holt, mai 2018 et Hillary Clinton says America is 'totally unprepared' for the impact of AI, de James Vincent, novembre 2017.

[1964] Voir Summary of the 2018 White House Summit on Artificial Intelligence for American Industry, mai 2018 (15 pages). Donald Trump n'était pas intervene dans ce sommet.

[1965] Voir DARPA announces $2B investment in AI, de Sarah Wells, septembre 2018 qui évoque le lancement du programme AI Next avec un budget de $2B étalé sur 5 ans. Les appels à projet concerneront la sécurité et la résilience des solutions d'IA, la réduction de la consommation d'énergie, les questions de performance et d'explicabilité. En plus de ce budget, il faut ajouter les $1,4B investis sur 5 ans dans le Joint Artificial Intelligence Center qui dépend du DSI du Département de la Défense.

[1966] L'enjeu du Mercury Challenge était modeste, avec un prix de seulement $100K, au total, pour 25 équipes. Le défi posé était de prédire des épidémies en Arabie Saoudite ou des révoltes pacifiques en Jordanie et en Egypte ainsi que des activités militaires en Syrie et en Iraq. Le tout avec des sources ouvertes d'information. Le concours se déroulait entre juillet 2018 et le printemps 2019.

[1967] Voir Trump's Plan to Keep America First in AI, février 2019. Les avis étaient partagés sur l'initiative de la Présidence, sachant que pour une fois, elle ne sortait pas de l'esprit tortueux du Président comme le mur et les tarifs douaniers, mais résultait de la consolidation apparente d'avis des agences fédérales en mode « bottom-up ». Voir What critics get wrong about the 'American AI Initiative' par Sujai Hajela, mars 2019 et The American AI Initiative: A good first step, of many par Mark Minevich, août 2019 et 4 Experts Respond to Trump's Executive Order on AI, The new "American AI Initiative" is heavy on bombast, light on specifics par By Eliza Strickland, février 2019.

Le décret présidentiel a été suivi par le dépôt d'une proposition de loi dénommée **Artificial Intelligence Initiative Act** introduite de manière bipartisane par des élus démocrates et républicains des deux chambres (version Sénat et version Chambre des Représentants). Cette loi proposait la création d'un Advisory Commitee sur l'IA, la création d'un Responsable de la coordination du plan (similaire au rôle de Bertrand Pailhès en France), l'allocation d'un budget pour la formation et la recherche, la création d'un comité de coordination inter-agences et le lancement d'une étude sur l'impact de l'IA sur l'emploi. Le plan était chiffré à $2,2B. Le conseil scientifique de la Maison Blanche publiait un rapport d'étape fin 2019[1968].

Début 2020, l'administration Trump décidait de renforcer ces investissements dans l'IA, toujours inquiétée par la montée en puissance de la Chine dans le domaine, tout comme dans le domaine des technologies quantiques[1969]. En gros, la proposition ensuite approuvée de manière bipartisane[1970] par le Congrès consistait à doubler l'investissement civil public dans l'IA de $1B à $2B sur deux années fiscales (2021 et 2022), sans que l'allocation soit précisées entre appels à projets NSF, subventions via les SBIR (projets financés dans le privé via les agences fédérales) et autres mécanismes de financements comme les dotations de laboratoires publics.

Diverses initiatives privées ou public-privées étaient aussi lancées avec notamment la création d'un accélérateur conjoint du **MIT** et de l'**US Air Force**, et celle par l'Université de **Stanford** d'un Institut de l'IA « for good »[1971]. On peut aussi remarquer côté fédéral la création du **Joint Artificial Intelligence Center** (JAIC) au Pentagone qui en est le centre d'expertise de l'IA et du projet **Maven** en 2017[1972]. La **DARPA** avait de son côté lancé une initiative sur l'IA de $2B en septembre 2018.

Du côté de la régulation, les USA doivent gérer un délicat équilibre avec d'un côté les lois antitrust qui pourraient amener l'Etat fédéral à limiter la puissance des GAFA et de l'autre, leur intérêt à préserver leur rôle dans l'exercice de la dominance sur l'Internet mondial. Certains candidats de la primaire démocrate de l'élection présidentielle de 2020 voulaient même démanteler les GAFA, surtout Elisabeth Warren. D'un point de vue pratique, l'Histoire montre qu'une majorité de ces tentatives dans le numérique ont conduit à réguler plus qu'à démanteler lesdits acteurs.

Sinon, dans la prolongation de sa stratégie de dérégulation, l'administration Trump avait prôné une approche diamétralement opposée aux approches européennes et françaises, limitant toute velléité de régulation de l'IA[1973]. On verra ce que l'administration Biden en fera !

Les USA ont aussi étendu des barrières à l'exportation de certaines technologies relatives à l'IA comme dans l'interprétation d'images satellite[1974].

Canada

Le gentil voisin des américains est aujourd'hui un des leaders de l'IA, tout du moins en comparaison avec le poids économique traditionnel du pays. La puissance publique n'y est pas étrangère, ayant été bien plus clairvoyante que ses homologues européens.

[1968] Voir 2016–2019 Progress Report Advancing Artificial Intelligence R&D, novembre 2019 (48 pages).

[1969] Voir White House proposes big increase in A.I. and quantum spending while cutting other sciences par Jonathan Vanian, février 2020. Cette proposition se faisait cependant au détriment d'autres domaines scientifiques comme ceux concernant l'environnement et... la santé, juste avant l'émergence de la crise épidémique du covid-19. Cependant, il faut regarder du côté du Congrès pour voir le fin mot de l'histoire car celui-ci ne suit généralement pas les propositions de l'administration Trump. Le financement des grands programmes de recherche a été ainsi sauvegardé. Voir Senate Releases FY21 Science Spending Proposals, novembre 2020.

[1970] Voir What's This? A Bipartisan Plan for AI and National Security par Tom Simonite dans Wired, juillet 2020.

[1971] Voir Stanford unveils new AI institute, built to create 'a better future for all humanity', mars 2019 et MIT And US Air Force Sign Agreement To Launch AI Accelerator, mai 2019.

[1972] Voir Project Maven Looks to Take Artificial Intelligence to Next Level, octobre 2017.

[1973] Voir New White House guidance downplays important AI harms par Alex Engler, décembre 2020.

[1974] Voir US announces AI software export restrictions par James Vincent, janvier 2020.

Les investissements en recherche en IA ont été préservés pendant le dernier hiver de l'IA tandis qu'ils diminuaient dans nombre d'autres pays. Les régions collaborent entre elles même si elles se concurrencent pour obtenir des fonds publics. Le pays a créé des instituts spécialisés dans l'IA en leur laissant une grande liberté dans les directions de recherche choisies. La proximité du marché US est aussi un avantage économique pour les startups. Cette stratégie a d'ailleurs aussi fonctionné pour le calcul quantique, autour des laboratoires de recherche et de la startup **D-Wave**.

La recherche canadienne s'appuie sur trois piliers : l'**Alberta Machine Intelligence Institute** (AMII) d'Edmonton, le **Vector Institute** à Toronto et le **MILA** de Montréal. C'est à Montréal que s'est installé l'un des laboratoires d'IA de **Facebook** (FAIR) dirigée par Joëlle Pineau ainsi que le laboratoire cortAIx de **Thales**[1975]. Les stars canadiennes de la recherche en IA sont **Geoff Hinton**, considéré comme le père du deep learning (maintenant chez Google), et **Yoshua Bengio** (toujours chercheur et ayant résisté aux sirènes des GAFAMI, au MILA de Montréal).

Le plan **Pan-Canadian Artificial Intelligence Strategy**[1976] de cinq ans lancé en 2017 était doté de 82M€. Ils sont investis dans la recherche, ce qui n'est pas grand-chose mais s'additionne à des financements existants. Les axes du plan étaient la recherche et la formation supérieure, la création de trois clusters scientifiques et le leadership sur les questions éthiques, économiques, légales de l'IA, un domaine qui a d'ailleurs donné lieu à un partenariat étroit avec la France. Le Canada a aussi lancé l'**Ivado** à Montréal en 2016, l'institut de valorisation des données, qui est un équivalent canadien de nos IRT (Instituts de Recherche Technologiques) associant chercheurs et industriels. Yoshua Bengio est le Directeur Scientifique de l'Ivado.

Cette excellence scientifique du Canada leur a permis d'y voir la création de nombreuses startups, l'une des plus connues était **Element.ai** (2016, $257M), malgré sa cession à ServiceNow fin 2020. Dans celles que nous avons déjà citées dans ce document, nous avons aussi **Lyrebird** (2017, $120K), **Mindbridge** (2015, $42,3M), **Interset** (2015, $24M), **LeddarTech** (2007, $123,3M), **Waste Robotics** (2016, $2,5M), **Analytics 4 Life** (2012, $29M), **Volpara Solution** (2005, $5,5M), **Deep Genomics** (2014, $16,7M), **Finn AI** (2014, $13,7M), **Ross Intelligence** (2014, $13,1M), **Kira Systems** (2015, $50M), **Twenty Billion Neurons** (2015, $12,5M), **Rubikloud** (2013, $45,5M), **Granify** (2011, $13,5M), **Landr** (2012, $34,9M) et **Nudge.ai** (2014, $4M).

Comme l'illustre le graphique *ci-dessous*, l'écosystème canadien de l'IA est dense, surtout comparativement à celui de la France[1977]. On y trouve les startups et les laboratoires de recherche. Le rapport associé fait état d'un tassement de la création de startups de l'IA en 2018 (+5% vs +28% en 2017) et une évolution naturelle de leur financement vers des tours de financement plus avancés. Ceux-ci ont représenté un total de $660M en 2018.

Le Canada a un point commun avec la France : un intérêt développé pour les questions d'éthique de l'IA. La **déclaration de Montréal** pour un IA responsable de fin 2018[1978] est signée par les chercheurs de référence canadiens que sont Yoshua Bengio, de l'Université de Montréal et Joëlle Pineau, de l'Université McGill et Facebook.

[1975] Voir Thales choisit le Canada pour son hub mondial en intelligence artificielle, octobre 2017. Basé à Montréal, le laboratoire cortAIx a été lancé en collaboration avec l'Institut Québécois d'Intelligence Artificielle (MILA), l'IVADO (Institute of Data Valorization), l'Institut d'Intelligence Artificielle du Québec et l'Institut Vector de Toronto.

[1976] Ce plan est piloté par le CIFAR, le Canadian Institute for Advanced Research, sorte d'équivalent du CNRS. Deux français font partie du conseil scientifique international de l'institut : Yann Le Cun (Facebook) et Antoine Petit (CNRS).

[1977] Source : Canadian AI Ecosystem 2019, par Jean-François Gagne, Peter Henderson et Yoan Mantha. On y découvre aussi que le Canada forme plus de spécialistes en IA qui s'expatrient que de spécialistes formés à l'extérieur du Canada et s'y installent. Même si cela porte sur une quantité faible de personnes : 370 vs 195 personnes.

[1978] Voir le site de la déclaration : https://www.declarationmontreal-iaresponsable.com/.

En décembre 2018, la France et le Canada lançaient un groupe d'experts internationaux sur l'IA, le **G2IA**, avec comme mission d'étudier notamment les dangers associés à l'IA[1979]. Ca a le goût d'un « machin » mais attendons pour voir. Le principal livrable de ce G2IA devait être une conférence internationale à organiser lors du G7 de Biarritz fin août 2019. En pratique, des discussions sur les recommandations du G2IA ont eu lieu entre les ministres du G7 en charge du numérique, en marge du sommet du G7.

Chine

La Chine a des ambitions de leadership mondial dans de nombreux secteurs de l'industrie et du numérique. L'IA ne fait pas exception à cette règle. L'Etat chinois a décidé de mettre le paquet sur l'IA et considère que c'est une technologie clé à maîtriser. Est-ce que pour autant les Chinois vont dépasser les GAFAMI américains ? C'est une hypothèse à la mode mais on en est train de se rendre compte qu'elle était un peu surfaite[1980].

Les grandes entreprises et startups chinoises du numérique bénéficient certes d'un terrain favorable : une masse de chercheurs et développeurs formés en Chine ou dans le reste du monde, un marché intérieur de plus de 900 millions d'Internautes (en 2020) et un marché mobile ultra-développé[1981]. Ces grands leaders chinois ont même recruté des talents chez leurs concurrents américains[1982]. La recherche chinoise fait sinon des progrès constants même si ils sont moins spectaculaires que ceux des chercheurs US/UK, comme ceux de DeepMind chez Google.

[1979] Voir Canada and France plan an international panel to assess AI's dangers par Will Knight dans MIT Technology Review, décembre 2018.

[1980] Voir China May Soon Surpass America on the Artificial Intelligence Battlefield, février 2017, China's AI awakening, paru en octobre 2017, Possible d'être un leader de l'AI d'ici 2030 ? La Chine en pôle position !, avril 2018, L'intelligence artificielle en Chine : un état des lieux par Aifang Ma, novembre 2018 (60 pages) ainsi que China AI Development Report 2018, China Institute for Science and Technology Policy, Tsinghua University, juillet 2018 (122 pages).

[1981] On compte notamment les BATX (Baidu, Alibaba, Tencent, Xiaomi) auxquels il faudrait au minimum ajouter Huaweï qui est le seul des grands chinois massivement présent hors de Chine en plus de Xiaomi. Et pour cause, c'est un fournisseur de technologies, pas un opérateur de services en ligne comme Baidu, Tencent ou Alibaba. Il y a aussi Wechat qui est aussi dans cette catégorie.

[1982] Rien que chez Baidu, Andrew Ng qui était auparavant chez Google tandis que Qi Lu provenait de Microsoft et Yahoo. Quand à Hugo Barra, ex Google passé avec fracas chez Xiaomi en 2013, il les a quittés début 2017 et est devenu VP de la réalité virtuelle chez Facebook. Ces prises restent cependant anecdotiques.

Enfin, les entreprises chinoises bénéficient d'une réglementation qui se pose beaucoup moins de questions sur la protection de la vie privée aussi bien dans la vidéo surveillance qu'en génomique. Cela se retrouve dans la généralisation à outrance de la vidéo surveillance.

On en découvrait un aspect peu réjouissant en 2020 avec cette évaluation en 2018 d'un système de reconnaissance faciale de Ouïghours développé par Huawei et Megvii, semble-t-il pas déployé[1983]. L'autre cas de figure est l'établissement d'un crédit social expérimenté dans plusieurs régions de Chine qui génère des frayeurs avec nos peaux d'occidentaux[1984].

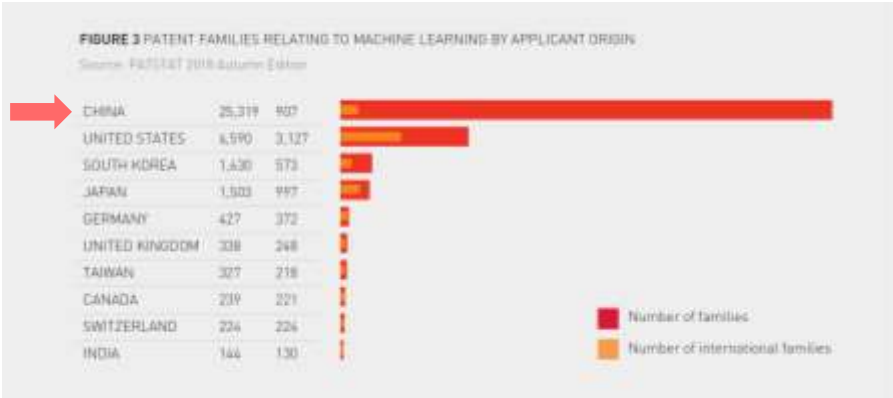
Les questions de vie privée ne sont pas importantes en Chine au regard de la gestion de « l'harmonie de la société » qui est plus importante que les libertés individuelles. Qu'on ne l'aime ou pas, c'est un référent culturel et politique différent des modèles occidentaux.

Le pays n'échappe cependant pas aux attentes de ses habitants de disposer d'un minimum de respect de leur vie privée. D'où l'éventualité de la création d'une loi de type RGPD qui protégerait les citoyens. Mais probablement plus vis à vis d'abus du secteur privé que du gouvernement[1985].

Petit à petit, la Chine acquiert son indépendance. On le voit dans les publications de chercheurs sur l'IA, notamment celles que l'on retrouve sur le site **arxiv** de Cornell University.

La Chine a largement dépassé les USA en termes de publications scientifiques dans l'IA (source).

Il en va de même pour les dépôts de brevets qui augmentent en flèche depuis 2015 et dépassent de loin des USA (ci-contre, par familles de brevets, en 2018[1986]).



source: Machine Learning Innovation A Patent Analytics Report Prepared by IP Australia, décembre 2019 (68 pages)

[1983] Voir Huawei tested facial recognition that targeted Uyghurs in China - The company acknowledged the report, but said it never deployed the system par Igor Bonifacic, décembre 2020. Voir aussi China's AI Unicorns Can Spot Faces. Now They Need New Tricks par Will Knight, décembre 2019 qui illustre les difficultés de Megvii à sortir de son champ initial de la reconnaissance faciale.

[1984] Voir The West may be wrong about China's social credit system par Bing Song, novembre 2018 qui illustre l'aspect très décentralisé et expérimental de ce crédit social. Et Intelligence artificielle : l'usage de l'IA par la Chine terrifie les experts par Bastien L, février 2019 qui évoque surtout la position de Yoshua Bengio. Un chercheur devient ainsi « les experts ». Voici un joli biais de la généralisation d'origine humaine.

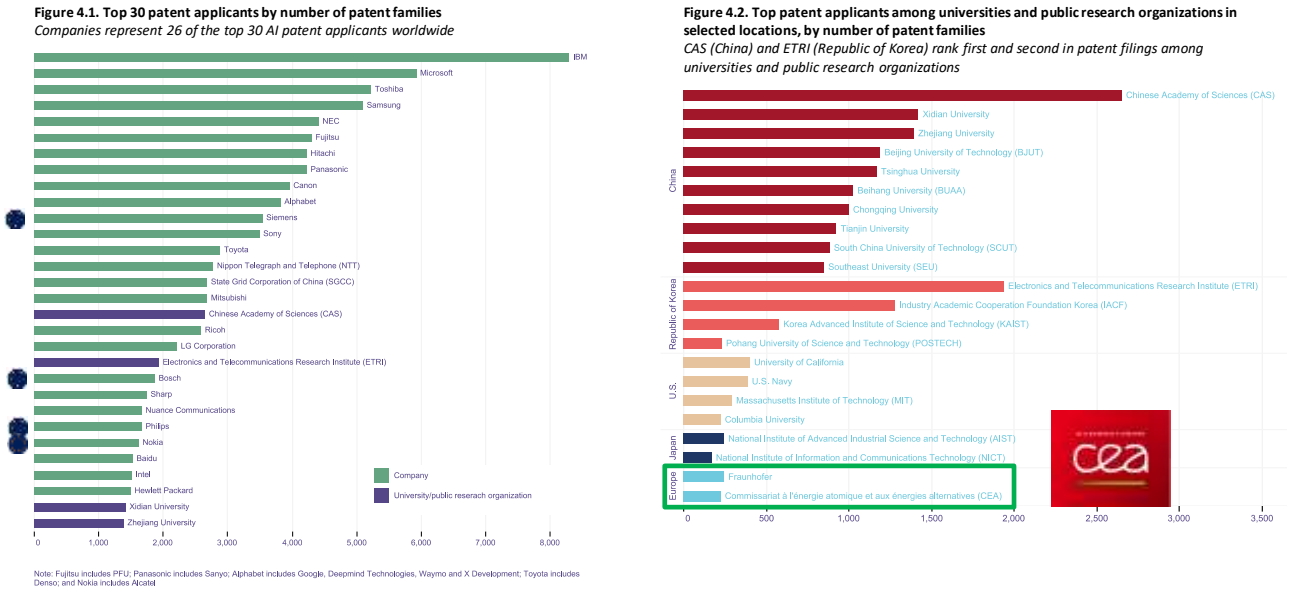
[1985] Voir Inside China's unexpected quest to protect data privacy par Karen Hao, août 2020 et La reconnaissance faciale angoisse les citoyens chinois selon une étude par Bastien L, décembre 2019.

[1986] Voir Machine Learning Innovation - A Patent Analytics Report, IP Australia, 2019 (68 pages).

Les pays occidentaux ne pourront en tout cas pas rattraper la Chine en faisant déposer des brevets pas ses IA stakhanovistes. En avril 2020, l'USPTO créait une jurisprudence en refusant d'accepter un dépôt de brevet au nom d'une IA, forçant les déposants à l'attribuer aux créateurs de l'IA[1987].

Dans le logiciel et chez les grands acteurs, la Chine affiche maintenant une stratégie agressive de protection et de valorisation de sa propriété intellectuelle. Cela ne concerne pas les mêmes intervenants dans l'industrie que ces PME chinoises du hardware qui ne se gênent pas pour donner dans la copie.

Cette dominance de la recherche chinoise en IA est remise en question par les chercheurs occidentaux[1988]. La nature des avancées chinoises est à relativiser selon les domaines[1989]. Ils sont visiblement meilleurs côté développement d'usages et applications. Ce qui en soi n'est pas une mauvaise chose puisque cela leur permet d'en tirer un profit économique plus substantiel.



source : WIPO Technology Trends 2019 Artificial Intelligence (158 pages).

Une analyse des brevets déposés dans le champ de l'IA par le WIPO en 2019 illustre une situation contrastée[1990]. Les déposants chinois sont surtout les universités tandis que les déposants américains et européens sont les grandes entreprises. En France, le CEA est le second déposant de brevets en Europe parmi les établissements de recherche tandis que le CNRS est le premier européen en publications scientifiques du domaine.

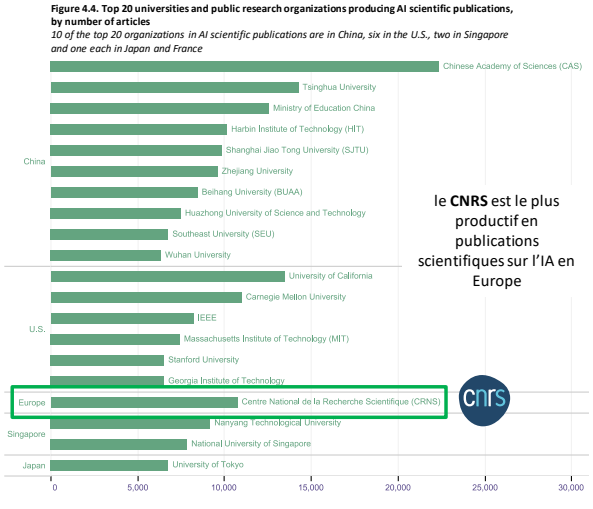
[1987] Voir USPTO Says AI Cannot Be Legally Credited As An Inventor par Tyler Lee, avril 2020.

[1988] Voir Is China really that far ahead in AI ? Research says no, de Frank Hersey, août 2018. Avec beaucoup de données sur la bataille des talents entre les USA et la Chine, montrant un avantage des USA, en tout cas en 2017. Voir aussi The State of Artificial Intelligence in China de Nanalyze qui fait référence au rapport China AI Development Report 2018 (122 pages) souligne notamment le fait que la proportion de top guns (h-index qui mesure l'impact des travaux des chercheurs via les citations de leurs papiers) parmi les chercheurs chinois est bien plus faible que dans les autres pays (5% vs 15% en moyenne). Et la France aurait plus de « top talents » que la Chine, tout comme le Royaume Uni, l'Allemagne et bien entendu, les USA. Mais c'est le seul indicateur rassurant parmi tous ceux que l'on peut trouver sur le poids de la Chine dans l'IA. J'ai découvert au passage **nr²** (2019, France) qui fournit des indicateurs intéressante et précis de l'état de l'écosystème d'innovation chinois. Voir aussi China's spending on AI may be far lower than people think par Meng Jing, 2019 et Who Is Winning the AI Race: China, the EU or the United States? par Daniel Castro, août 2019.

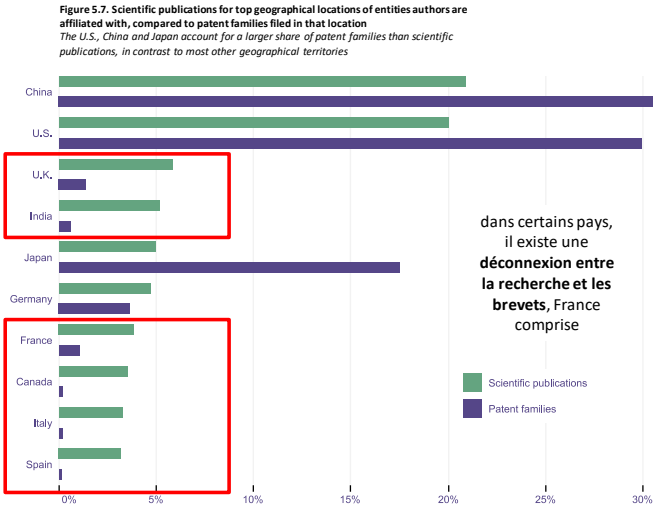
[1989] Voir What you may not understand about China's AI scene par Karen Hao, MIT Technology Review, avril 2020, qui évoque le fait que l'on exagère la performance chinoise dans l'IA mais souligne le fait que les chercheurs chinois sont avantagés car ils sont nombreux à pouvoir absorber les travaux occidentaux publiés en anglais tandis que les chercheurs occidentaux ne peuvent pas lire les publications chinoises.

[1990] Voir WIPO Technology Trends 2019 Artificial Intelligence (158 pages).

Par contre, la France, le Royaume Uni, l'Inde, le Canada, l'Italie et l'Espagne se distinguent par une faible transformation des publications en brevets, témoignant d'une difficulté à valoriser économiquement le fruit de la recherche tandis que les USA et la Chine ont un très bon ratio brevets/publications.



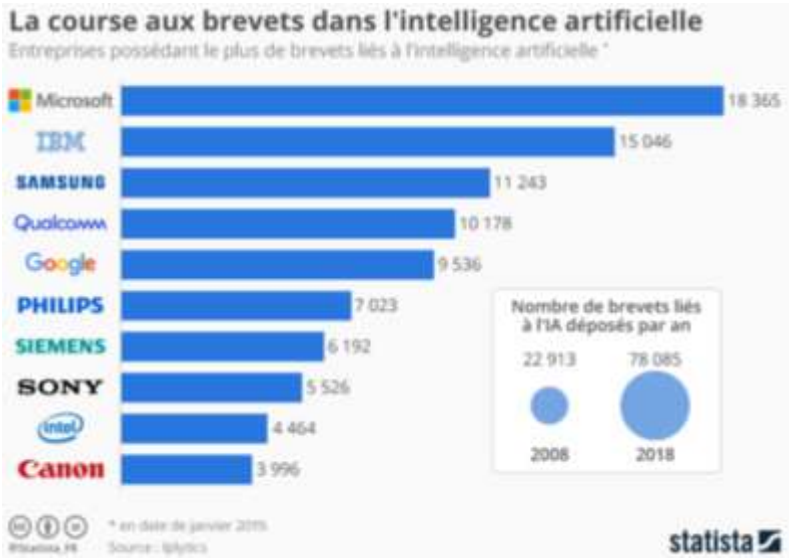
le **CNRS** est le plus productif en publications scientifiques sur l'IA en Europe



dans certains pays, il existe une **déconnexion entre la recherche et les brevets**, France comprise

Malgré tout, il subsistera encore longtemps un cloisonnement des marchés au niveau des données et des applications entre la Chine et le monde occidental.

Sauf accidents de parcours, les grands chinois ne feront pas l'acquisition de GAFAMI. Par contre, en termes de grands groupes déposant en brevets sur l'IA, les sociétés américaines dominent clairement le jeu avec quelques européennes (Siemens, Philips) et japonaise (Canon)[1991].



Si **Baidu** et d'autres font quelques progrès significatifs dans le deep learning[1992], les techniques utilisées qui sont généralement open source ne sont pas des différenciateurs stratégiques significatifs par rapport à ce qui se fait en Occident.

Seule la donnée acquise dans leur pays l'est et les données captées par les leaders chinois manquent de diversité pour bien couvrir les besoins à l'échelle mondiale[1993]. La majeure partie des grands acteurs de l'Internet chinois n'ont pas de présence dans les pays occidentaux.

[1991] Voir La course aux brevets dans l'intelligence artificielle par Tristan Gaudiaut, mai 2019.

[1992] Voir Baidu spins out its global ad business to sharpen its focus on artificial intelligence de Jon Russel, mai 2018.

[1993] Si l'on s'intéresse à la reconnaissance de visages, Google pourrait avoir un avantage sur les leaders chinois. En effet, la qualité d'un tel système est liée à la diversité des visages qui servent à son entraînement. Google est mieux placé pour disposer d'une grande diversité de visages dans ses bases. Il en va de même pour Facebook, ne serait-ce que de par la grande diversité des visages que l'on peut trouver aux USA. Cette diversité est bien moindre en Chine. Pour bien entraîner une IA à base de deep learning, la variété des données fournies est clé ! On pourrait appliquer le même raisonnement à plein d'autres types de données qui sont dépendants de la langue et de la culture : la parole, les modes de consommation, les modes de vie, de transports, etc. Bon, ceci dit, SenseTime est une licorne chinoise qui bat tous les records et sa spécialité est la reconnaissance de visages pour la vidéosurveillance.

Wechat, Tencent, Weibo ou Douban sont inconnus chez nous à ce stade de leur développement. Seul Alibaba concurrence dans une moindre mesure Amazon[1994]. Baidu voudrait aussi s'y essayer, ayant décidé de déployer pas moins de 5 millions de « serveurs cloud intelligents » d'ici 2030 et de former 5 millions de professionnels de l'IA en 5 ans.

Par contre, sur certains marchés et notamment dans des technologies enfouies, les Chinois peuvent adopter une stratégie mondiale. C'est le cas de **Baidu** avec son système d'exploitation Appolo OS destiné aux véhicules, notamment autonomes et **Huawei** avec ses processeurs mobiles Kirin et ses infrastructures pour les opérateurs télécoms.

Les grands acteurs et les startups sont en effet très hardis dans la conception de chipsets d'IA, même si les avancées réelles qu'ils réalisent ne sont pas évidentes par rapport à l'état de l'art occidental, notamment de Nvidia[1995] ou de startups comme Cerebras (USA) ou Graphcore (UK).

Les startups chinoises sont très focalisées sur la vision artificielle, notamment pour les applications de vidéo surveillance comme avec **SenseTime**, **Yitu**, **CloudWalk** et **Megvii**[1996]. Elles ciblent surtout leur marché intérieur.

La Chine a par contre une forte déficience dans l'influence du monde des développeurs dans l'IA. Très peu de frameworks et outils de développement utilisés en Occident proviennent de Chine aux exceptions près de ceux de Baidu (DuerOS et Apollo OS pour l'automobile). Bref, comparer la Chine aux USA en matière d'IA n'est pas évident. Cela dépend des indicateurs utilisés.

Du côté du gouvernement, la Chine a lancé un **Next Generation Artificial Intelligence Development Plan** en 2016 et annoncé publiquement en juillet 2017[1997]. Ce plan aurait été provoqué par le choc de la victoire d'AlphaGo contre Ke Jie, le champion du monde Chinois du jeu, en mai 2017. Un moment « Spoutnik » d'après Kai Fu Lee. Mais la prise de conscience de l'importance avait visiblement démarré bien avant cet événement[1998].

Le plan du gouvernement couvre la R&D, l'industrialisation et notamment dans le domaine des chipsets, la formation et la capacité à attirer les meilleurs talents mondiaux, l'influence dans la définition de standards et régulations, dans la création de normes éthiques et pour la sécurité.

Le plan devait se dérouler en trois phases : un rattrapage là où doit y en avoir d'ici 2020, le leadership dans certains domaines d'ici 2025 et le leadership général d'ici 2030. Ils ambitionnent de créer une industrie de l'IA de $144B d'ici 2030 et le décuple dans son impact sur les autres industries, dont $1,4T. Pour mémoire, le PIB de la Chine en 2017 était de $12,24T. Ces chiffres contiennent les mêmes biais que certaines prévisions de McKinsey sur les marchés de l'IA et des technologies quantiques, confondant le marché des technologies d'IA et celui des clients qui y font appel.

Côté thématiques, la Chine se focalise sur les systèmes intelligents de type véhicules autonomes, robots de services, l'automatisation des usines, les systèmes d'authentification et de vidéosurveillance ainsi que dans la santé. Côté hardware, ils visent les capteurs intelligents et les chipsets neuromorphiques.

[1994] Voir A map of the Chinese social media & internet ecosystem, mai 2018 et China Internet Report 2018 (97 pages).

[1995] Voir New Chips, L4 Autonomous Bus & Baidu Brain 3.0 Showcased at Baidu Create 2018, juillet 2018 et Alibaba to launch own AI chip to avoid overdependence on US, septembre 2018. L'activité de chipsets d'Alibaba est liée à l'acquisition du Chinois C-Sky Microsystems en avril 2018.

[1996] Voir Deep Learning Startups in China : report from the leading edge, juillet 2017 ainsi que Beijing subways may soon get facial recognition and hand scanners, juin 2018.

[1997] Voir A Next Generation Artificial Intelligence Development Plan, juillet 2017 et plus de détails dans China's AI Agenda Advances d'Elia Kania, février 2018.

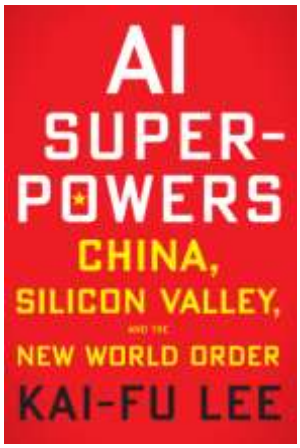
[1998] Ceci fait suite à des initiatives favorisant l'entrepreneuriat de masse en 2014 avec la multiplication d'incubateurs et accélérateurs dans les grandes villes, encouragés par l'Etat et financés par les collectivités locales. Il y aurait à la mi 2018 environ 6600 incubateurs de startups en Chine. Les collectivités locales ont aussi créé leurs fonds d'investissements. Le capital risque privé s'est enfin développé par la même occasion.

Ils souhaitent en particulier ne pas dépendre des USA sur ce dernier point, surtout au vu de la bataille économique qui fait rage sur les droits de douane et les interdictions d'exportations ciblées qui touchent directement ou indirectement les entreprises chinoises comme Huawei[1999].

Le plan comprenait le développement d'un parc technologique de recherche en IA de $2,1B AI à Beijing, sur 55 hectares pour attirer 400 entreprises dans les domaines des réseaux très haut débit, du cloud, des systèmes biométriques et du deep learning. Ils prévoyaient d'y installer la 5G et un supercalculateur. Si la Chine regorge de chercheurs d'excellents niveaux, elle met aussi le paquet dans la formation d'ingénieurs en IA qui vont faire adopter cette dernière par l'ensemble des industries. Le déploiement de l'IA dans des villes intelligentes pilotes est aussi prévu[2000].

Ce centre de recherche en IA a été créé et inauguré en février 2018, dénommé **Beijing Frontier International AI Research Institute**. Il fournit des ressources en données et en calcul aux startups en IA chinoises, notamment celles qui sont focalisées sur la vidéosurveillance. Rien qu'à Beijing, il y a plus de 400 startups dont 160 dans l'IA et qui peuvent tirer parti de ces nouvelles ressources.

Ce centre est dirigé par Kai-Fu Lee qui était successivement chez Apple, Microsoft (MSR China) puis Google. Il avait créé en 2009 le fonds d'investissement **Sinovation Ventures** dédié aux startups en Chine.



Il est aussi l'auteur de « AI super powers China, Silicon Valley and the new world order » en 2018. L'ouvrage décrit l'émergence de l'entrepreneuriat Internet en Chine et notamment dans la mobilité. Il illustre comment les startups américaines se sont trompées en abordant le marché chinois, en n'adaptant pas assez leurs offres et applications aux spécificités du marché chinois. Il montre comment les startups chinoises, via une mécanique concurrentielle inouïe et darwinienne, ont pu reprendre le contrôle de leur marché intérieur. Kai Fu Lee grossit le trait de la supériorité de l'Internet chinois avec le concept O2O, pour « online-to-offline », les applications mobiles populaires comme WeChat reliant les utilisateurs mobiles aux services et activités commerçantes du monde physique.

La différence chinoise est la concentration de ces activités dans des plateformes mobiles généralistes comme WeChat ou la plateforme d'achats groupés Meituan-Dianping, sorte de Groupon local, qui est notamment devenue très populaire pour la livraison de repas ainsi que pour la location de vélos, via son acquisition de Mobike en 2018.

Nombre de ces entreprises n'hésitent pas à gérer la partie physique de leur business alors que leurs homologues américains restent plutôt cantonnés à l'immatériel. Le Uber local, Didi, a une valorisation boursière supérieure à celle d'Uber depuis début 2018 (mais ce sont des valorisations d'entreprises non cotées) et opère ses propres stations services et garages. De même, le Airbnb local, Tujia, gère ses propres lieux d'hébergement.

Kai Fu Lee met en avant le fait que la Chine a plus d'utilisateurs que l'Amérique du Nord et l'Europe réunies. Mais il néglige le décalage de PIB par habitant ($9K/an en Chine, $59K/an aux USA et autour de $42K/an en France/Allemagne/UK). L'approche internationale de ces mêmes startups s'appuie sur l'aide ou l'investissement dans des startups locales, notamment dans les pays émergents, plutôt que sur des déploiements mondiaux à l'américaine[2001].

[1999] Ces interdictions ont eu plusieurs impacts : impossibilité de faire fabriquer des semiconducteurs chez TSMC à Taiwan car celui-ci fait appel à des équipements d'origine américaine comme ceux d'Applied Materials, l'impossibilité d'intégrer la version certifiée Android de Google ou des restrictions sur l'utilisation de noyaux arm. Et c'était avant que Nvidia mette la main sur arm !

[2000] Voir China's AI City to put computers in charge par Umberto Bacchi, décembre 2020.

[2001] Voir The Chinese and American Apps Winning the Next Billion Users par Matt Sheehan, avril 2020 qui montre qu'à part en Inde, peu d'application grand public chinoises arrivent à s'imposer dans le monde, face à celles qui proviennent des USA.

L'ouvrage de Kai-Fu Lee décrit l'émergence de l'entrepreneuriat Internet en Chine et notamment dans la mobilité. Il illustre comment les startups américaines se sont trompées en abordant le marché chinois, en n'adaptant pas assez leurs offres et applications aux spécificités du marché chinois.

Bref, il montre comment les startups chinoises, via une mécanique concurrentielle inouïe et darwinienne, ont pu reprendre le contrôle de leur marché intérieur.

Kai Fu Lee grossit le trait de la supériorité de l'Internet chinois par le concept O2O, pour *online-to-offline*, les applications mobiles populaires comme WeChat reliant les utilisateurs mobiles aux services et activités commerçantes du monde physique. La différence chinoise est dans la concentration de ces activités dans des plateformes mobiles généralistes comme WeChat ou la plateforme d'achats groupés Meituan-Dianping, sorte de Groupon local, qui est notamment devenue très populaire pour la livraison de repas ainsi que pour la location de vélos, via son acquisition de Mobike en 2018.

Kai Fu Lee explique aussi pourquoi les USA se font distancer par la Chine[2002]. Il considère que les USA n'investissent pas assez dans des projets ambitieux de recherche pour faire avancer radicalement la portée de l'IA. Il décrit la pénurie de talents alimentée aux USA par d'un côté les GAFAMI et de l'autre une politique migratoire répressive[2003], ce en quoi il n'a pas tort.

Il décrit un futur sombre généré par l'IA, déstabilisant l'équilibre social et économique du monde, augmentant les inégalités et le chômage. Au passage, on ne parle pas de GAFA et de BATX en Chine pour ce qui est de l'IA, mais des 7 géants que sont Google, Facebook, Amazon, Microsoft, Baidu, Alibaba, et Tencent. Donc trois Chinois sur 7 géants ! Il manque juste Huawei pour égaliser. Le livre de Kai Fu Lee évoque enfin les risques que l'IA fait peser sur l'emploi à partir de sa moitié avec une digression sur le sens de sa vie et de l'amour.

Le projet *ci-dessous* est-il issu d'un laboratoire de recherche chinois ? Non, il provient d'une Université d'Albany aux USA[2004]. C'est un scénario que le gouvernement chinois souhaiterait éviter. Autant il apprécie d'attirer en Chine des talents mondiaux, autant il est agacé lors que des Chinois s'expatrient à l'étranger et y restent pour faire de la recherche ou créer des startups.

C'est un souci qu'ils ont depuis plus de 10 ans. L'État chinois en est au point où il souhaite limiter les échanges d'étudiants et d'ingénieurs avec les USA. Le schéma de droite illustre ce problème, les étudiants chinois qui se forment aux USA y restant en grande majorité.

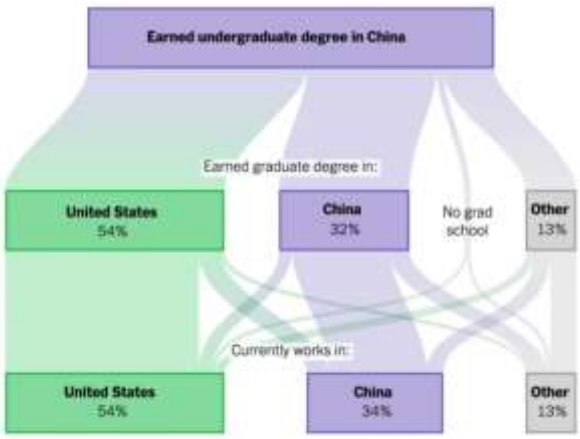
Le marché chinois est sinon toujours aussi protectionniste. Autant les grands acteurs chinois de l'Internet sont absents des marchés occidentaux à de rares exceptions, autant il est difficile pour les leaders américains de s'implanter en Chine. Facebook, YouTube et Twitter n'y sont pas et y sont même interdits. Seul Google reviendrait à la charge avec son moteur de recherche qui respecterait les contraintes du gouvernement chinois.

Voir Faut-il avoir peur du plan de la Chine pour dominer l'intelligence artificielle? par Philippe Silberzahn, février 2020

[2002] Voir The US is hastening its own decline in AI, says a top Chinese investor, octobre 2018.

[2003] Voir Intelligence Artificielle : quand la Chine aura pris le pouvoir... par Louis Naugès, novembre 2018, qui reprend à son compte les thèses de Kai Fu Lee. Un peu trop rapidement car il ne décrit pas les modalités de la supériorité chinoise.

[20042004] J'ai trouvé ce projet dans la présentation Video Analytics for AI City Smart Transportation, 2017 (41 slides). Voir China's path to AI domination has a problem: brain drain par Karen Hao, août 2019 qui fait reference à China's AI Talent Base Is Growing, and then Leaving de Joy Dantong Ma, juillet 2019 qui fait état du fait que quasiment les trois quarts des chercheurs chinois en IA exercent hors de Chine. Sachant que l'étude est basée sur les contributeurs de la conférence internationale NeurIPS.



Royaume Uni

L'approche anglaise est intermédiaire entre celle des USA et de la France pour ce qui est du rôle des pouvoirs publics, avec un interventionnisme mesuré compensé par un entrepreneuriat et un financement des startups plus dynamique.

Le pays se distingue depuis longtemps comme étant le plus dynamique côté startups, aussi bien pour leur création que pour leur financement. Et dans l'IA comme ailleurs. Loin devant l'Allemagne et la France, ces dernières espérant que le Brexit portera une entaille sur cette dominance. Il en va de même des capitales, Londres devançant Paris et Berlin dans son dynamisme entrepreneurial, en tout cas, mesuré en nombre de startups et surtout, en levées de fonds.

Londres bénéficie également du plus grand nombre de salariés de filiales de groupes américains (Microsoft, IBM, Oracle, Facebook, Google). Ceux-ci le sont pour au moins deux raisons : le chiffre d'affaire de ces entreprises au Royaume-Uni a toujours été bien plus élevé qu'en Allemagne et en France, le Royaume Uni étant un très gros consommateur d'informatique et de numérique ramené au PIB et au nombre d'habitants, et ensuite, le pays étant anglophone, il bénéficie d'une proximité culturelle et politique plus forte avec les USA.

Côté planification, cela a commencé avec le rapport Growing the Artificial Intelligence Industry in the UK de Wendy Hall et Jérôme Pesenti, en octobre 2017 (78 pages). Il s'agissait d'un rapport indépendant associant une chercheuse et un entrepreneur, en l'occurrence, un Français, ce qui est fort étonnant mais s'explique par la position de Jérôme Pesenti à l'époque[2005]. Le rapport préconisait diverses mesures dans l'ouverture des données, dans le développement des compétences, dans l'organisation de la recherche en IA structurée autour d'un UK AI Council, la modernisation de l'Etat avec l'IA. Le Royaume Uni avait déjà créé en 2015 le Alan Turing Institute qui devient des facto le point de coordination de la recherche en IA dans le pays avec un maillage dans les universités et laboratoires de recherche publics.

[2005] Doctorat en math et en philosophie, Jérôme Pesenti était à l'époque CEO de **Benevolent.ai** (2013, UK, $207M) qui associait l'IA à la médecine pour extraire le savoir des publications scientifiques afin de créer de nouveaux traitements en biotech. Auparavant, il avait créé la startup **Vivisimo** (2000, USA, $5,7M) qui était spécialisée dans l'extraction de données. Après son acquisition par IBM en 2012, il dirige l'équipe technique d'**IBM Watson** dont un tier des effectifs provient de Visimo. Après son passage par Benevolent.ai et depuis début 2018, il est le patron de l'IA chez **Facebook** et devient le patron de Yann Le Cun qui gère les laboratoires FAIR (Facebook AI Research). Il partage avec ce dernier le fait d'être un optimiste de l'IA.

S'en suivait un autre rapport, produit par la Chambre des Lords, AI in the UK ready, willing and able ?, 2017 (183 pages)[2006] et pour lequel Jérôme Pesenti était d'ailleurs auditionné. Dans ce rapport, la France n'était pas du tout citée comme étant un pays investissant significativement dans l'IA. Pour l'Europe, seule l'Allemagne était citée et bien sûr, le Canada, les USA et la Chine. Ce rapport était un peu l'équivalent de celui de l'OPECST français de mars 2017. Il faisait le point sur les questions éthiques, sociétales et économiques induites par l'IA. Il s'inquiétait notamment des monopoles qui peuvent se constituer autour des données. Il recommandait l'audit des données et d'accompagner les PME dans l'adoption de l'IA via un fonds. Il souhaitait faire du Royaume-Uni le champion mondial de l'IA éthique !

En avril 2018, le gouvernement et les parties prenantes de la recherche et des entreprises de l'IA publiait Artificial Intelligence Sector Deal (21 pages). Il était cosigné par deux Ministres, celui en charge de l'industrie et l'autre, en charge du numérique et qui avait aussi la culture, les médias et le sport dans son portefeuille, ainsi que par la chercheuse Wendy Hall et Jérôme Pesenti, alors entré chez Facebook. Le plan mettait en regard les investissements publics et ceux des entreprises. Il insistait sur l'investissement au Royaume-Uni d'entreprises étrangères de l'IA venant des USA, du Japon, du Canada et de Hong Kong : Google, Element AI, Amazon, HPE, Beyond Limits, Ironfly, Astroscale et Chrysalix. Les investissements publics spécifiques à l'IA étaient de £618M et le privé y met au moins £300M.

Etait également créé le **Centre for Data Ethics and Innovation** pour traiter des questions d'éthique de l'IA qui lançait en juin 2018 une consultation publique[2007]. En mai 2018, Theresa May annonçait un investissement de £210M dans l'IA et la santé, pour permettre aux chercheurs et entrepreneurs d'exploiter les données accumulées par le National Heath Service dans la lutte contre le cancer[2008].

De manière concommitante ou redondante, la NHS lançait début 2019 un plan d'investissement sur 10 ans de $330M. Il démarrait en 2020 avec un appel à projets piloté par NHSx, la branche technologique du système de santé public du Royaume-Uni, doté de $66M pour sa première phase. L'idée est de favoriser la collaboration entre l'état, les acteurs du système de santé, les chercheurs et les entreprises pour le déploiement de solutions à base d'IA[2009].

Enfin, en juin 2018, le gouvernement UK répondait point par point aux propositions du rapport de la Chambre des Lords d'avril 2018[2010]. Le gouvernement insistait beaucoup sur la collaboration avec le privé. Il intervient de manière ponctuelle sans que les investissements soient gigantesques.

[2006] Voir UK can lead the way on ethical AI, says Lords Committee, avril 2018.

[2007] Voir Centre for Data Ethics and Innovation Consultation, juin 2018.

[2008] Voir Britain to Invest Heavily in AI for Medicine, dans MedGadget, mai 2018.

[2009] Voir U.K. Invests $330 Million To Lead The World In Healthcare AI par Dr. James Somauroo, décembre 2020.

[2010] Voir Government responds to report by Lords Select Committee on Artificial Intelligence, juin 2018 et la réponse : Government response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able? (42 pages).

La forte présence d'entreprises étrangères est un atout pour le gouvernement. Ils l'intègrent d'ailleurs dans la communication de **TechNation**, leur équivalent de le French Tech depuis 2017.

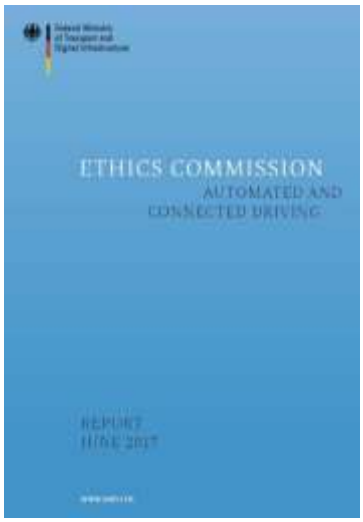
L'approche du Royaume-Uni est donc un deal entre le public et le privé, ce dernier intégrant même explicitement les entreprises étrangères du numériques, surtout américaines, établies dans le pays.

A noter une conséquence du Brexit gênante pour le Royaume Uni : la décision d'**IBM** de transférer en Allemagne ses clusters GPU d'entraînement Watson, histoire de respecter les contraintes du RGPD[2011].

Allemagne

L'Allemagne est assez discrète sur l'IA en tant que telle. Son industrie est mobilisée dans l'Industrie 4.0 qui est à l'industrie allemande ce que la transformation digitale est aux entreprises française, à savoir un grand fourre-tout. Un second domaine est prioritaire : l'industrie automobile.

Le gouvernement présentait les premiers éléments de sa stratégie dans l'IA en juillet 2018. Il souhaite accélérer les transferts technologiques entre la recherche et l'industrie. La recherche en IA est concentrée dans le German Research Centre for AI (DFKI), la Alexander von Humbolt Foundation et les Fraunhofer Institute. Il insiste sur la collaboration européenne et notamment avec la France. On y trouve les éléments habituels comme la formation, l'attraction de talents étrangers, la modernisation des services publics avec l'IA et la promotion d'une IA éthique. Sur ce dernier point, le gouvernement annonçait la création d'une commission dédiée à l'éthique de l'IA avec 19 députés et 19 experts de l'IA. Elle doit créer un set de recommandations d'ici 2020. Le plan IA allemand est doté de 3Md€ étalés sur cinq ans entre 2019 et 2024[2012].



Sur un modèle équivalent, le Ministère des Transports et de l'Infrastructure Numérique avait publié un rapport en juin 2017 sur l'éthique des véhicules autonomes[2013]. Il traitait des dilemmes dans la prise de décision sur le choix de sauver telles ou telles vies humaines. La réponse ? Il n'y a pas de solution universelle. Ils indiquent que les véhicules autonomes doivent tout faire pour éviter toute forme d'accident. Pour la commission, la morale ne peut pas être intégrée dans les systèmes de conduite. Ils recommandent même en cas d'accident d'éviter de faire quelque discrimination positive ou négative en fonction de l'âge, du genre, de l'état physique des personnes concernées par les dilemmes. Par contre, il serait justifié de minimiser le nombre total de victimes d'un accident donné mais les parties prenantes à l'origine de risques ne doivent pas sacrifier des parties tierces qui n'y sont pour rien.

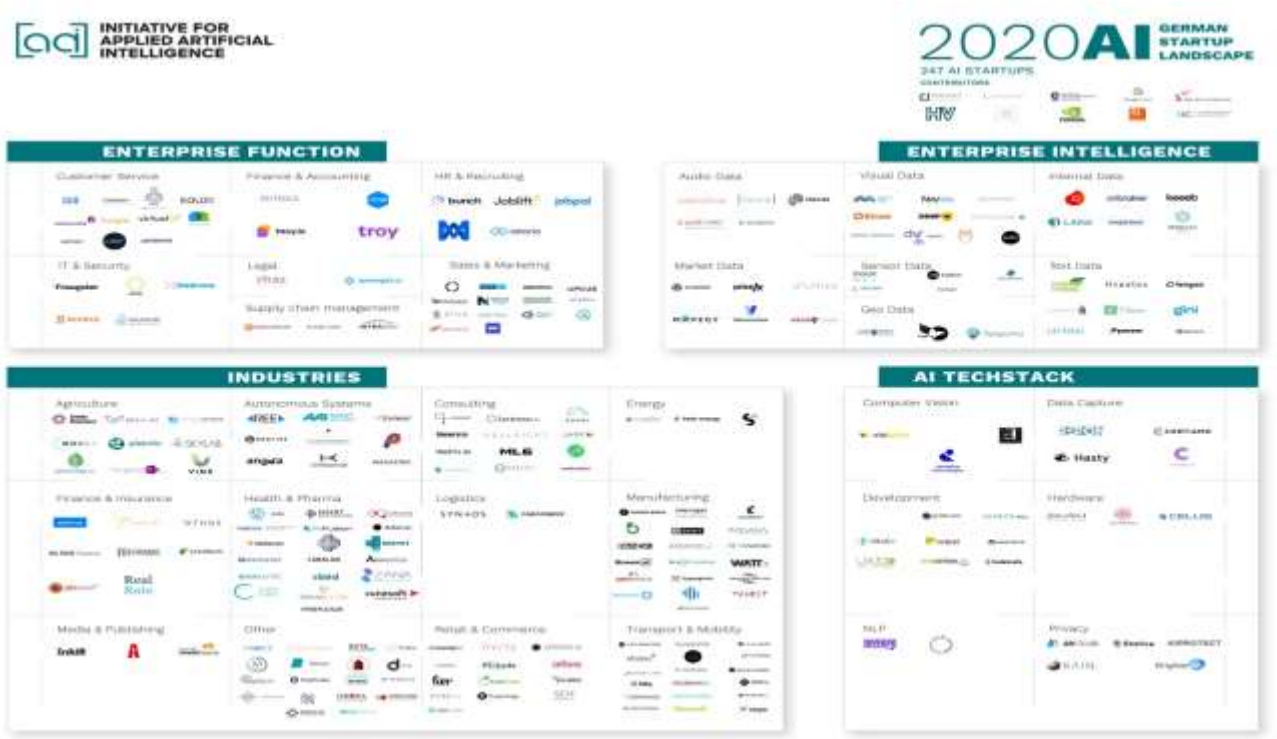
Ils recommandent une mesure voisine de ce qui se fait dans l'aviation : la création d'un bureau d'analyse des accidents impliquant des véhicules autonomes pour apprendre des accidents et ainsi améliorer les systèmes de conduite autonome.

En novembre 2018, le gouvernement allemand lançait une véritable stratégie nationale sur l'IA en lançant un investissement fédéral de 3Md€ abondés par autant issus du secteur privé. Le plan se focalisait sur la recherche avec le financement d'une centaine de chaires universitaires, de douze centres de recherche en IA complétant les laboratoires existant comme le DFKI (German Research Centre for Artificial Intelligence).

[2011] Voir Watson Machine Learning: GPU Cluster Migration from London to Frankfurt par IBM, avril 2020.

[2012] Voir Germany AI Strategy Report.

[2013] Voir Ethics Commission Automated and Connected Driving Report 2017 (36 pages).



La situation n'est pas bien brillante côté startups de l'IA, tout du moins si l'on s'en tient aux levées de fonds[2014].

En janvier 2021, les plus grandes levées de fonds de startups dans l'IA comprenaient celles de **Kreditech** ($519M), **Wefox** ($268M), **Konux** ($130M), **Freeletics** ($70M), **Ada Health** ($69M), **Acrolinx** ($60M), **Arago** ($60M), **Zeitgold** (51,2M€), **Savedroid** ($53M), **Door2door** (31,8M€), **Fineway** ($25M), **Dojo Madness** ($18,8M) et **Leverton** ($18M), sachant que dans le tas, la moitié relèvent de l'IA washing. Cela ne fera donc pas de l'Allemagne un leader mondial de l'IA, tout du moins par le biais de ses startups. Un inventaire de 2020 fait état de 247 startups dans le secteur de l'IA, en retrait par rapport à la France. Ces startups sont très orientées sur le marché des entreprises et de l'industrie, ce qui n'est pas étonnant.

Estonie

L'Estonie est connue pour être le pays le plus numérisé d'Europe, notamment au niveau de ses services publics construits autour d'un bus de données unique. En mai 2019, ils lançaient un plan de modernisation des services publics à base d'IA dénommé Kratt[2015]. Ils investissent notamment dans les usages de l'IA dans le système de santé.

Du côté des startups, Nanalyze inventorie huit startups de l'IA ayant levé entre $200K et $11,2M. Elles sont positionnées dans des secteurs variés : éducation, sécurité, finance, RH, relation client et divers[2016]. C'est une position marginale, comme celle de tous les pays de l'Union Européenne.

Union Européenne

En avril 2018, la Commission Européenne publiait une première ébauche de plan sur l'IA[2017]. Les objectifs affichés étaient de développer la capacité industrielle européenne dans l'IA dans le privé et le public.

[2014] Voir Top-10 Artificial Intelligence Startups in Germany, dans CBInsights, mai 2018.

[2015] Le plan Eesti tehisintellekti est détaillé mais rédigé en estonien, ce qui n'est pas bien pratique, mai 2019 (65 pages).

[2016] Voir 9 Artificial Intelligence Startups in Estonia, avril 2019.

[2017] Voir Communication Artificial Intelligence for Europe, avril 2018 (20 pages),

Il s'agissait aussi de se préparer aux bouleversements sociétaux et économiques induits par l'IA et de développer un cadre éthique de l'usage de l'IA. Un investissement de 500M€ était prévu en 2017 et d'un total de 1,5Md€ d'ici 2020[2018].

C'était suivi en juin 2018, par la création d'un groupe de réflexion sur l'IA, le **HLEG AI** qui coordonne les débats de l'**European AI Alliance**, un forum de « policy making »[2019]. Le groupe est le résultat d'un savant dosage (par pays, genre et métiers). Il comprend 52 personnes de l'industrie, de la recherche et de la société civile dont neuf Français[2020].

Ce groupe de travail publiait des recommandations sur l'éthique de l'IA en avril 2019 comprenant un ensemble de sept propositions pour une IA digne de confiance[2021].

Il enchaînait en juin 2019 avec un rapport comprenant 33 groupes de recommandations sur un grand nombre d'aspects [2022]: la création d'IAs de confiance, centrées sur l'humain, de la formation et de l'éducation du public sur le sujet[2023], de mieux rémunérer les chercheurs et faciliter les transferts de technologie vers l'industrie avec des hackathons et des partenariats public-privé, de faire levier sur les Etats clients, et d'être les champions de l'IA éthique en interdisant tout système de crédit social à la chinoise et en ne mettant pas en place de système de surveillance de masse en général.

Les experts y dénoncent surtout le déficit d'investissements dans l'IA comparativement aux USA et à la Chine. Ils demandent une aide publique des Etats et de l'Union Européenne. Mais comme c'est souvent le cas dans ce genre de rapport, les propositions opérationnelles sont vagues, surtout d'un point de vue quantitatif. Les questions de défense y sont abordées uniquement sous l'angle de la cybersécurité[2024].

L'Union Européenne souhaite aussi favoriser les partenariats transnationaux qui conditionnent souvent les financements, notamment ceux d'Horizon 2020 et son successeur **Horizon Europe** qui couvrira la période 2021-2027[2025] ou des *flagship projects*. Ces projets transnationaux sont poussés par les pays de toutes tailles et par les chasseurs professionnels de subventions. Le problème est que ces projets génèrent beaucoup de friction managériale sans que ce soit forcément justifié par leur taille. Les projets européens ont du sens lorsqu'il s'agit d'atteindre une taille critique comme ce fut le cas avec Airbus, Amadeus ou le CERN.

[2018] Voir With €1.5 billion for artificial intelligence research, Europe pins hopes on ethics de Tania Rabesandratana, avril 2018.

[2019] Voir High-Level Expert Group on Artificial Intelligence.

[2020] Les neuf représentants français sont Yann Bonnet (ex CNNum et rédaction du rapport de la Mission Villani), Nozha Boujemaa (Inria), Raja Chatila (Sorbonne), Mari-Noëlle Jégo-Laveissière (Orange), Raoult Mallart (Sigfox), Françoise Soulié Fogelman (consultante), Thierry Tingaud (STMicroelectronics), Cécile Wendling (Axa) et Thiébaut Weber (secrétaire général de la confédération européenne des syndicats). Donc, aucune startup, à part Sigfox.

[2021] Voir Ethics guidelines for trustworthy AI, avril 2019 (56 pages). L'un de ces experts s'est désolidarisé du rapport. Voir IA : « l'éthique-washing », une invention européenne par Thomas Metzinger, 2019. Il considère qu'une IA digne de confiance est un récit marketing de l'industrie pour endormir les clients et que les propositions du groupe d'experts sont des demi-mesures insuffisantes.

[2022] Le rapport Policy and investment recommendations for trustworthy Artificial Intelligence, juin 2019 (52 pages) et L'effort européen pour l'IA est « insuffisant », et c'est la Commission qui le dit par Rémy Demichelis, juin 2019.

[2023] Qui peut s'inspirer d'une initiative finlandaise consistant à former en ligne la population. Voir Comment la Finlande a formé 1% de sa population à l'intelligence artificielle par Benoît Georges, 2019.

[2024] Voir en complément Artificial Intelligence – What implications for EU security and defence? par Daniel Fiott et Gustav Lindstrom, tous deux de l'EU Institute for Security Studies (EUISS) qui est basé à Paris, novembre 2018 (8 pages).

[2025] Voir Horizon Europe (57 slides).

En février 2020, l'Union Européenne publiait un livre blanc censé définir une stratégie sur l'IA[2026]. En guise de stratégie, le propos portait plus sur le besoin d'éviter de créer le far-west et donc, de réguler. On ne peut pas dire qu'il s'agisse d'une véritable vision industrielle. Toutes ces visions qui visent à organiser ou régimenter les données sont vaines car elles négligent les services qui en sont à l'origine. Le problème de la poule et de l'oeuf entre services et données dans l'IA n'est pas bien compris. La dominance américaine du secteur côté infrastructures comme logiciels est trop focalisée sur les données alors qu'elle relève de facteurs anciens qui favorisent les acteurs américains qui bénéficient de meilleures économies d'échelle que les acteurs européens qui agissent dans un marché hyper-fragmenté. Résultat, l'IA est abordée comme un champ d'usage devant être adopté par les entreprises plus que comme une industrie en tant que telle. Cette dichotomie est habituelle dans le numérique depuis longtemps[2027].

En matière d'offre, l'Union Européenne dispose toutefois d'instruments financiers. Europe Next qui remplace Europe 2020 finance la R&D en amont. La BEI finance les entreprises, notamment par des prêts.

Elle gère l'EIF (European Investment Fund) qui lançait en 2019 un fonds d'investissement de 2Md€ couvrant notamment l'IA et la blockchain[2028]. En pratique, il s'agit d'un maximum de 400M€ pour ces deux domaines, 100M€ devant venir de la BEI et le reste d'investisseurs privés. Comme quoi il faut toujours se méfier des effets d'annonces !

Les initiatives ne manquent pas comme ce projet de création d'un institut européen de l'IA sur le modèle du CERN, l'**Ellis Institute** (European Lab for Learning and Intelligent Systems)[2029]. Ce n'est pour l'instant qu'une proposition de scientifiques, qui vont ensuite s'étriper pour choisir le ou les pays où implanter cet institut s'il est approuvé et financé. La France et l'Allemagne seraient sur les rangs[2030]. Il est un secteur où de tels projets pourraient avoir du sens : les semiconducteurs. C'est l'objet du projet **EPI** (European Processor Initiative) visant à créer un processeur européen pour les supercalculateurs mais qui pourrait aussi servir aux véhicules autonomes, une demande des industriels allemands.

Il ne sera pas facile à concilier les deux objectifs ! Ce projet regroupe 23 partenaires issus de 10 pays vise à assurer une indépendance européenne face aux USA. Le projet est dirigé par Philippe Notton d'**Atos** et ancien de STMicroelectronics. Ce processeur à basse consommation doit servir à créer des supercalculateurs exascale.

Mais il ne s'agit visiblement pas d'une architecture de tenseur adaptée aux applications de l'IA. Le projet sera « fabless », la fabrication du processeur pouvant être réalisée chez Infineon en Allemagne, STMicroelectronics à Grenoble où en Asie s'il faut descendre en-dessous de 28 nm. Le projet sera financé par l'Europe à hauteur de 120M€[2031].

[2026] Voir On Artificial Intelligence - A European approach to excellence and trust, février 2020 (27 pages).

[2027] Voir Europe and AI: Leading, Lagging Behind, or Carving Its Own Way? par Erik Brattberg et al, 2020, L'écosystème d'excellence proposé par l'Europe pour l'IA : utopie ou réalité ? par Bob De Caux, (date pas trouvée), Intelligence artificielle: cinq choses à savoir sur les projets de l'UE par l'AFP, février 2020 et Thierry Breton : « La guerre des données industrielles débute maintenant et l'Europe sera son principal champ de bataille » par Derek Perrotte, février 2020.

[2028] Voir EU launches €2bn AI and blockchain fund | Sifted par Maija Palmer, novembre 2019.

[2029] Voir Scientists plan huge European AI hub to compete with US de Ian Sample, avril 2018.

[2030] La recherche française signataire de cette lettre comprend Marc Schoenauer (Inria, et corapporteur du rapport de la Mission Villani) et Stéphane Mallat (Collège de France), Cordelia Schmid (Irnia Grenoble), Francis Bach (Irnia Paris), Jean Ponce (Inria), Julien Mairal (Inria), Jakob Verbeek (Inria), Jean-Philippe Vert (Mines ParisTech et ENS Paris), Emmanuel Dupoux (EHESS, ENS, Paris Sciences Lettre, CNRS, Inria), Theodoros Evgeniou (Insead), Karteek Alahari (Inria), Alexandre d'Aspremont (CNRS, ENS Paris), Pierre-Paul Zalio (ENS Paris-Saclay). On y trouve aussi des Anglais, des Israéliens, des Américains et des Suisses !

[2031] Voir European Processor Initiative: consortium to develop Europe's microprocessors for future supercomputers, mars 2018.

Le projet est un sous-ensemble de l'EuroHPC (European High-Performance Computing) qui vise à créer un supercalculateur européen. Reste à savoir s'il pourra être réellement européen. Quid de la mémoire et du stockage ? Il risque au minimum d'être asiatique.

Du côté entrepreneurial, l'Europe pâtit toujours de la faiblesse de ses acteurs industriels « pure players » du numérique. Les investissements dans l'IA en Europe étaient situés entre 2,4Md€ et 3,2 Md€ en Europe pour une fourchette de 12 à 18Md€ en Amérique du Nord et de 6,5 à 9,7 Md€ en Asie[2032].

Mais l'Europe est-elle perdue pour ce qui est de l'IA comme l'affirme de manière péremptoire Kai Fu-Lee et avec des arguments à l'emporte-pièce [2033]? Non. Si l'Europe a bien du mal à faire germer des acteurs locaux et régionaux couvrant les besoins du grand public, elle arrive tout de même à être présente dans les offres d'entreprises. Et surtout lorsqu'elles exportent rapidement aux USA. C'est le cas dans l'Histoire française de Dassault Systèmes, Business Objects, Talend, Docker, Datadog, Neolane ou Criteo.

Il existe de nombreuses associations professionnelles européennes autour de l'IA. Il y a notamment **EurAi**, l'European Association for Artificial Intelligence (anciennement ECCAI) a été crée en 1982. Elle organise la conférence ECAI tous les deux ans en Europe. Et puis **euRObotics** qui regroupe surtout les industriels européens de la robotique ainsi qu'**AI4EU**, qui avait été notamment valorisée lors de l'**European AI Night**, un événement organisé à Paris en avril 2019 au Palais de Tokyo.

Israël

Israël est un OVNI de la sphère mondiale des startups avec une concentration de startups sans égale ramenée au PIB et au nombre d'habitants. Nous n'allons pas refaire l'histoire de l'écosystème entrepreneurial de ce pays qui est très bien décrite dans l'ouvrage **Startup Nation** de Dan Senor et Saul Singer (2009).

Le pays cumule tous les facteurs de succès des startups : culturels (éducation, ambition, international, polyglotte), économiques et politiques (proximité avec les USA, pas de marché intérieur).

Résultat, dans l'IA, les startups du pays sont abondantes. L'inventaire Israel Artificial Intelligence Startups de 2018 comprenait 950 startups ayant levé en tout $7,5B en septembre 2018. 85% d'entre elles sont b2b. Ceci est à comparer avec environ 500 startups françaises de l'IA financées aux alentours de 1,4Md€ selon France Digitale dans une étude d'octobre 2019 déjà citée. A noter que, rien que dans la vision artificielle, l'inventaire Israel Computer Vision Startups publié en mars 2019 fait état de 245 startups financées à hauteur de $2,9B ! Voilà le benchmark de référence pour l'Europe !

Notons que le pays est à l'origine de la plus grande sortie en date dans le domaine de l'IA, avec l'acquisition de **Mobileye** par Intel en mars 2017 pour $15,3B, le spécialiste de la vision artificielle embarquée dans les véhicules.

Russie

Vladimir Poutine a contribué à politiser la course à l'IA en septembre 2017 en déclarant que le pays qui deviendrait leader du domaine deviendrait le dirigeant du monde.

[2032] Selon McKinsey, dans 10 imperatives for Europe in the age of AI and automation, 2017.

[2033] Voir Kai-fu Lee: there is no hope for Europe's artificial intelligence sector par Carly Minsky, décembre 2018. Et Europe is losing the AI race par Michael Stothard, janvier 2019 qui se base sur les indicateurs de dépôts de brevets en oubliant peut-être qu'en Europe, on a moins l'habitude de déposer des brevets sur le logiciel qui sont en théorie interdits par les accords de Munich sur la propriété intellectuelle. Et lorsque Kai-Fu Lee affirme qu'il n'existe pas d'écosystème de VC et d'entrepreneurs en Europe, il se paye notre tête ! Dans "*Plus, Europe has no VC-entrepreneur ecosystem. The entrepreneurs in Europe are nowhere near as innovative as the American ones nor as tenacious as the Chinese ones. Europe's entrepreneurs lack the experience of dealing with these types of software and artificial intelligence problems*".

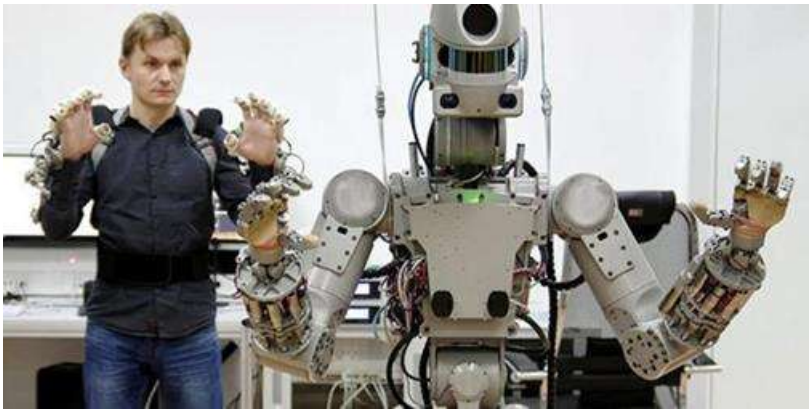
Il précisait qu'il ne serait pas souhaitable qu'un pays s'arroge un tel contrôle et qu'il faudrait partager le butin. C'est ce que l'on dit lorsque l'on a peu de chances de le récupérer.

En effet, c'est un moyen d'admettre que la Russie serait imanquablement sous la houlette des Chinois ou des USA ! Les chances que la Russie joue un rôle de premier plan mondial dans l'IA sont assez maigres au vu de leur production dans la recherche et de leur rythme de création de startups. La vision de Poutine est manichéenne et simpliste.

Les investissements publics de la Russie dans l'IA ne seraient que de $12,5M par an, une bien maigre bouchée de pain[2034] ! Leurs milliardaires préfèrent s'acheter des yachts que de financer la formation des jeunes à l'informatique comme le fait Xavier Niel en France avec l'école 42.

Une bonne partie des investissements russes dans l'IA sont militaires. C'est même le Ministère de la Défense qui donne le la, dans sa liste des 10 priorités publiée en mars 2018[2035].

On y trouve du big data, de la formation à l'IA, la création d'un centre national de recherche en IA et l'organisation de « war games » à base d'IA. D'où le fait que leurs prouesses dans l'IA sont souvent militaires ou voisines des questions militaires : un système de guidage autonome pour avion de chasse SU-35 ou un robot manutentionnaire qui sera embarqué dans l'ISS d'ici 2020/2021[2036].



Du côté des startups, on constate en effet que les leaders russes n'ont pas fait de mirifiques levées de fonds, témoignant au minimum de leur faible impact international [2037] . Comme on le voit ci-contre, aucune levée supérieure à $10M ce qui témoigne d'une difficulté à « scaler ».

Leur marché intérieur est insuffisant pour se développer. Le PIB par habitant de la Russie est en effet du niveau de celui du Brésil donc bien au-dessus de celui de la France et du Royaume-Uni, avec des fluctuations dépendant des cours des matières premières.

[2034] C'est un montant étonnament faible. En voici la source : In AI, Russia Is Hustling to Catch Up, avril 2018.

[2035] Vu dans Conférence "Intelligence artificielle: problèmes et solutions" – 2018, traduit en français via Google, mars 2018.

[2036] Voir Russia's sudden change of heart on AI par Dov Zakheim, mai 2019.

[2037] Voir The Top-10 Russian Artificial Intelligence Startups, juin 2018.

En octobre 2019, Vladimir Poutine lançait une nouvelle stratégie nationale de développement de l'IA. L'idée est d'assurer le développement accéléré de l'IA dans le pays d'ici 2030, avec en tête l'obtention de l'indépendance technologique et avec une forte connotation militaire et cyberdéfense (ou cyberattaque, mais cele ne s'avoue pas)[2038]. Le plan associé comprend des financements de la R&D, de l'open data publique et de la formation des compétences. Le budget de tout cela restait à déterminer ! La Russie dispose déjà d'un fonds souverain de $2B financé par des investisseurs privés d'Asie et du Moyen-Orient.

Le danger provenant de la Russie est surtout géopolitique. Ses services de renseignement n'hésitent en effet pas du tout à employer des moyens techniques asymétriques pour faire une nouvelle guerre froide aux démocraties occidentales. Leur ingérence dans l'élection présidentielle US de 2016 via la combinaison des services actifs du GRU et de la web farm Internet Research Agency financée par un oligarque russe a été dénoncée par plusieurs mises en examen du procureur spécial Robert Mueller en 2018[2039].

Les techniques employées ne faisaient cependant pas visiblement pas appel à de l'IA contrairement au machine learning mis en œuvre par Cambridge Analytica pour profiler et cibler les électeurs américains. Et rien ne dit que les cyberattaques SunBurst de fin 2020 se sont appuyées sur la détection de vulnérabilité à base d'IA du logiciel ciblé, Orion de l'Américain SolarWinds.

Corée du Sud

La Corée du Sud a connu un choc lorsque DeepMind AlphaGo a battu le champion du monde du jeu de Go, le Corée Lee Sedol, en 2016. Deux jours après cette victoire très médiatisée, le gouvernement coréen annonçait un investissement de 770M€ dans la recherche en IA sur quatre ans[2040]. Ce plan implique évidemment les grands groupes industriels coréens à commencer par Samsung, LG, Hyundai et aussi Naver. Hyundai est évidemment intéressé par l'IA pour le développement de véhicules autonomes. Samsung a déjà mis sur le marché sont assistant personnel Bixby depuis début 2017. Et Naver investit dans les startups via sa mise de 100M€ dans le fonds d'investissement français Korelya créé en 2016 par Fleur Pellerin.

L'autre approche de la Corée consiste à installer des laboratoires de recherche à l'étranger. Les entreprises américaines font cela pour des raisons de lobbying et pour prétendre être bien intégrées économiquement avec les grands pays où ils exportent.

Ici, il s'agit véritablement de capter des compétences locales. Cela explique l'installation d'un laboratoire de recherche de **Samsung** à New York en robotique ou celui de **Naver** à Grenoble, qui y a repris les équipes de recherche de Xerox.

En 2018, le gouvernement coréen lançait un nouveau plan de $1,5B, toujours focalisé sur la R&D. Comme partout ailleurs, le plan vise à accélérer la formation de talents dans l'IA (5000 étudiants, 1400 chercheurs en IA et 3600 data scientists), développement de technologies d'IA dans la santé, la défense et la sécurité, le tout en s'appuyant sur un challenge voisin de ceux qui sont lancés par la DARPA.

Le dernier volet est un plan pour développer des semiconducteurs pour l'IA à commencer par un projet d'étude de $28M[2041].

[2038] Voir La stratégie russe de développement de l'intelligence artificielle par Thierry Berthier et Yannick Harrel, décembre 2019.

[2039] Voir l'inculpation sur la web farm de février 2018 (37 pages) et celle concernant le GRU, juillet 2018 (29 pages).

[2040] Voir South Korea trumpets $860-million AI fund after AlphaGo 'shock', mars 2016

[2041] Voir South Korean Government to Invest Billions of Dollars into Developing AI Semiconductors, février 2018. D'ailleurs, il n'est pas nécessaire d'investir des milliards de dollars pour créer des composants d'IA. En effet, la Corée dispose déjà des capacités de fabrication de semiconducteurs de Samsung. Les besoins en R&D sont donc surtout fabless et peuvent se contenter de quelques dizaines de millions de dollars.

Pour créer un chipset d'ici 2029, ce qui semble être une bien lointaine échéance à moins qu'il ne s'agisse de créer une véritable rupture technologique comme avec des memristors, qu'il est très difficile de mettre au point.

Taïwan

Avec ses 24 millions d'habitants, Taïwan a une influence technologique quasiment aussi grande que son voisin la Chine continentale. Une bonne partie des grandes entreprises d'assemblage (Hon-Hai/Foxconn, Quanta…) sont taïwanaises même si la majorité de leurs usines d'assemblage sont situées en Chine continentale.

Taïwan comprend aussi Acer, Asus, Gigabyte, MSI, HTC et surtout le fondeur TSMC qui est l'un des trois restants dans le monde dans la course à l'intégration en-dessous de 10 nm avec Samsung et Intel, et encore, ce dernier est mal en point de ce point de vue-là. TSMC joue un rôle critique dans le hardware de l'IA puisque c'est lui qui fabrique une bonne part des ASIC du marché : ceux de Nvidia, Apple, Graphcore, Cerebras, Tesla, AMD, Qualcomm et plein d'autres encore.

Le pays investit beaucoup en R&D. Ses PME sont plutôt innovantes en comparaison de leurs homologues chinoises, surtout celles de Shenzhen. Elles exportent toutes bien évidemment. Taïwan est un peu l'Israël de l'Asie, en concurrence avec Singapour et la Corée du Sud.

Le gouvernement de l'île a annoncé quelques initiatives autour de l'IA. En novembre 2018, il lançait l'installation d'un supercalculateur d'IA construit par Quanta Computers, Asus et Taiwan Mobile, Taiwania 2. Le supercalculateur comprend 252 unités de calcul dotées de deux Intel Xeon Gold et 8 Nivida V100 totalisant 9 PFLOPS. C'est un projet voisin du supercalculateur du CNRS déployé en France par HPE. Le pays lançait aussi l'initiative TWCC (Taiwan Compute Cloud).

En parallèle, Taiwan lançait une initiative sur les composants de l'IA, l'AITA (**AI on Chip Taiwan Alliance**) qui rassemble 56 entreprises du secteur. L'enjeu ? La course aux chipsets de l'IA pour serveurs et l'embarqué. L'idée est de faire en sorte que Taiwan ne soit pas que l'usine de fabrication de ces chipsets (Qualcomm et Nvidia y font appel) mais aussi un concepteur de ces chipsets de plus en plus stratégiques. L'AITA intègre également un objectif de création d'outils de développement logiciels pour ces chipsets. Les membres de cette initiative comprennent TSMC, UMC (un autre fondeur), MediaTek, Realtek Semiconductor, Nanya Technology, Quanta Computer, Foxconn Electronics, Asustek Computer et Microsoft Taiwan.

Du côté des startups, c'est toujours Nanalyze qui fournit un inventaire intéressant des plus grandes levées de fonds[2042]. On y distingue quelques startups significatives comme **Appier** (2012, Taiwan, $81M, dans le marketing en ligne et la publicité), **Viscovery** (2013, Taiwan, $15M dans la vision artificielle, notamment pour le retail et le paiement automatique) et **iKala** (2011, Taiwan, $13M, commerce en ligne).

Japon

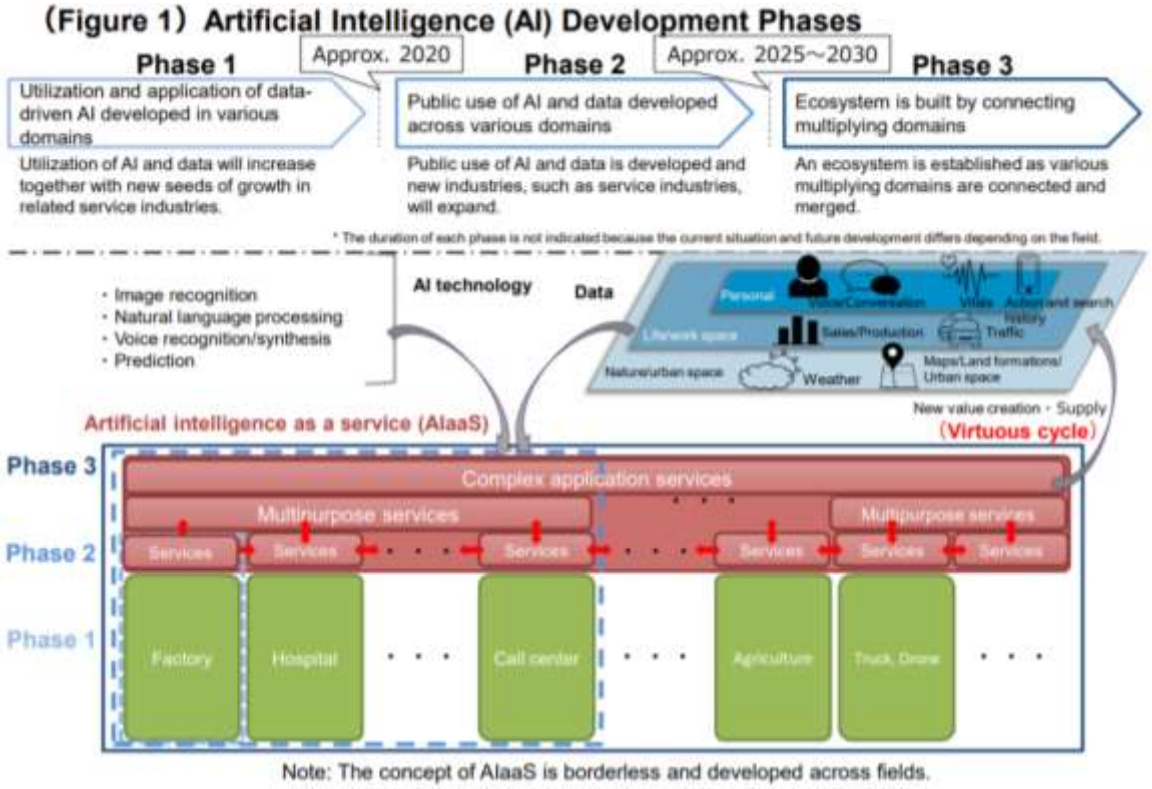
Le pays a été un des premiers à formaliser une stratégie dans l'IA, en avril 2016, sous la forme d'un plan pluriannuel structuré à l'ancienne comme les Japonais en ont le secret (*ci-dessous*). Cela comprend le développement de technologies de base puis d'applications verticales dans les marchés habituels de l'IA (industrie, santé, agriculture…)[2043]. Ses trois phases sont l'application de l'IA dans divers marchés, l'usage d'IA et de données dans des usages grand public, et la création d'un écosystème multi-discipline. Cela reste très théorique. Il leur reste aussi à dynamiser leur écosystème entrepreneurial, qui est faiblement développé au regard de ceux que l'on trouve en Amérique du Nord, en Europe et même en Chine.

[2042] Voir Top-9 Artificial Intelligence Startups in Taiwan, mai 2019.

[2043] Le plan japonais est formalisé dans Artificial Intelligence Technology Strategy, mars 2017 (25 pages).

Le gouvernement japonais décidait en 2019 d'allouer l'équivalent de $921M sur 10 ans pour financer des projets en robotique avec, toujours en ligne de mire, l'idée toujours repoussée aux Calendes Grecques, de créer des robots de services pour les personnes âgées, pour s'adapter à l'inexorable vieillissement de la population[2044].

De son côté, **Softbank** mettait la barre un peu plus haut en annonçant la création d'un fonds d'investissement sur l'IA de $108B en juillet 2019, Vision Fund 2[2045]. Ce fonds Vision Fund 2 a une ambition mondiale, pas seulement japonaise.



C'est aussi un moyen de mettre la main sur des pépites occidentales, comme Softbank l'avait fait avec Aldebaran Robotics et Boston Dymanics en robotique[2046].

[2044] Voir Le Japon investit 1 milliard dans la recherche sur l'homme augmenté et les cyborgs par Nathan Le Gohlisse, août 2019.

[2045] Le premier fonds Vision Fund, créé en 2016 et abondé par l'Arabie Saoudite, avait déjà été doté de $100B mais n'investissait pas spécifiquement dans l'IA. En 2019, ce fonds était présent dans 24 des 377 licornes mondiales identifiées. Ce genre de fonds est complexe à analyser : les sources de financement sont très variées entre fonds souverains, milliardaires, grandes entreprises mondiales, etc. Leur rôle dans les stratégies des pays concernés n'est pas évident.

[2046] Voir SoftBank announces AI-focused second $108 billion Vision Fund with LPs including Microsoft, Apple and Foxconn, juillet 2019.

Du côté des startups, on en trouve quelques-unes de significatives, tout du moins côté levée de fonds. Elles ne sont pas forcément bien connues chez nous[2047]. La raison est un syndrome courant dans l'innovation logicielle : de nombreuses startups se lancent sur les mêmes créneaux un peu partout dans le monde et exploitent les mêmes technologies, et elles sont souvent des sociétés de service déguisées avec un marché plutôt de proximité. A noter Leapmind, créateur de blocs fonctionnels de processeurs d'IA.

Name	Application	City	Funding *(USD millions)*
Preferred Networks	Internet of Things	Tokyo	130.0
ABEJA	Big data	Tokyo	45.4
Ascent Robotics	Autonomous vehicles	Tokyo	17.9
Cinnamon	Optical character recognition (OCR)	Tokyo	17.0
LeapMind	Edge devices	Tokyo	13.5
Cogent Labs	OCR, natural language understanding	Tokyo	12.9
Moneytree	Fintech	Tokyo	10.5
MUJIN Inc.	Robotics	Tokyo	7.0
Alpaca	Fintech	Tokyo	6.7
MJI	Smart assistant	Tokyo	5.0

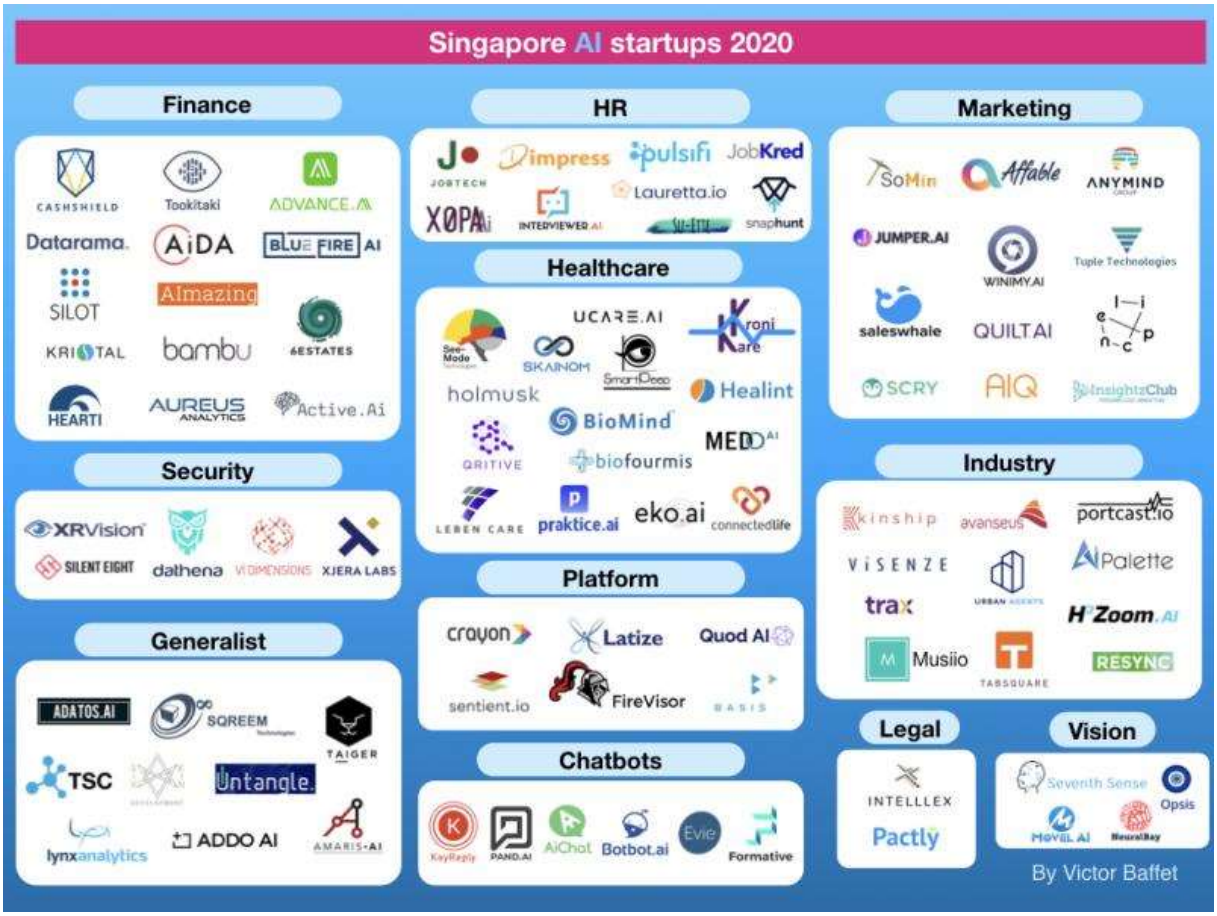
Singapour

L'écosystème des startups de Singapour est assez vivace en intelligence artificielle comme d'ailleurs, en technologies quantiques. Le pays bénéficie d'une position géographique intéressante, proche de la Chine et au centre de la Malaisie. On y trouve notamment une belle licorne, **Trax Technology** (2010, Singapour, $386M) qui est spécialisée dans la vision artificielle pour le retail. Leur solution analyse ce qui se passe dans les rayons des supermarchés pour optimiser les ventes. Ils sont présents aux USA. Le montant levé est énorme.

Comme dans de nombreux pays développés, l'écosystème de startups en IA est riche et diversifié[2048]. Il est complété depuis 2020 par un « hub mondial de l'innovation » spécialisé dans l'IA que l'Allemand **Infineon** veut établir sur place... d'ici 2023 et pour un investissement de $27M. Il s'agit visiblement comme le font d'autres tels qu'IBM d'un centre d'avant-vente et de support technique de projets utilisant les technologies d'Infineon. 1000 personnes doivent être formées sur place pour s'adapter à cette nouvelle mission.

[2047] Voir Top-10 Artificial Intelligence Startups in Japan, février 2019.

[2048] Cette cartographie édition 2020 a été crée par Victor Baffet.



Inde

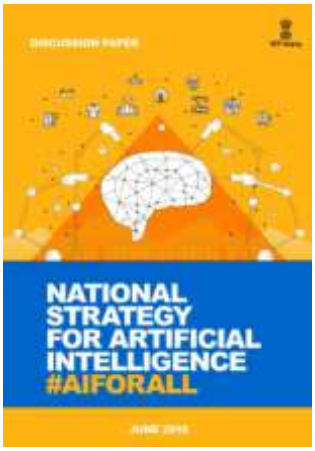
L'Inde est le troisième pays du monde en termes de publications scientifiques dans l'IA mais le cinquièment en nombre de citations[2049]. C'est aussi le second pays du monde en termes de population. Mais il est beaucoup moins développé que la Chine. Le PIB par habitant en Inde est de $1939 pour $8826 en Chine, un rapport de 1 à 4,5 ! Le pays forme 2,6 millions d'étudiants dans les sciences par an, mais trop dans l'informatique traditionnelle et pas assez en IA.

Le pays a publié sa stratégie dans l'IA en juin 2018[2050]. Elle focalise le pays sur ses besoins intérieurs : dans la santé notamment dans les outils de diagostic, dans l'agriculture de précision, dans l'éducation, la smart city et les transports. S'y ajoute le traitement du langage pour la traduction, très utile pour jouer le rôle de passerelle entre les nombreuses langues parlées en Inde, en plus de l'anglais.

[2049] Mais il a bien du mal à transformer cela en innovations et en startups.

[2050] Voir National Strategy for Artificial Intelligence, juin 2018 (115 pages). C'est un plan très bien documenté et qui s'appuie beaucoup sur la description de bonnes pratiques internationales, dont la DARPA aux USA et PRAIRIE en France.

La situation en Inde est bien différente de la France du côté du marché de l'emploi. En 2016, 51% des emplois étaient dans l'agriculture. Le pays prévoit que cela va descendre à 37% d'ici 2030, soit le niveau d'après guerre en France. Ils prévoient que ce trou d'air sera en grande partie compensé par les emplois dans la construction et dans une moindre mesure dans la santé.



Leur activité d'offshore est impactée depuis une douzaine d'années par le déploiement de solutions de RPA (Robotic Process Automation) qui automatise des processus de backoffice de sociétés dans la banque et l'assurance qui faisaient de l'offshore de backoffice en Inde. Il leur faut compenser cela avec des emplois plus qualifiés, d'où les efforts en formation.

Du côté des startups, leur écosystème comprend beaucoup de copycats de projets américains comme **Manthan** ($98M, big data), **Sigtuple** ($25M, vision artificielle), **Haptix** ($12M, chatbot), **Rubique** ($10M, finance), **CreditVidya** ($7M, finance). L'économie numérique du pays dépend surtout des activités de services incarnées par les géants tels que **Wipro** et **Infosys** tout deux situés à Bengalore, et qui pourront développer leurs activités de gestion de projets d'IA b2b qui auront besoin d'autant de prestations de services qu'avant[2051]. Une situation qui fait penser à celle de la France avec ses nombreuses entreprises de services numériques !

Emirats Arabes Unis

Ce petit ensemble de 7 émirats totalisant 9,7 millions d'habitants posé sur un désert en bord de mer n'a pas la prétention de devenir un pôle de recherche et industriel de l'IA. Leur ambition est d'être un utilisateur exemplaire de l'IA dans différents secteurs du gouvernement et de l'économie locale : transports, santé, éducation, espace, énergies renouvelables, eau et environnement.

A l'occasion de la présentation de leur plan stratégique sur l'IA[2052], les Emirats Arabes Unis sont les premiers à avoir créé un Ministère de l'Intelligence Artificielle en octobre 2017 dont le premier titulaire, Omar Sultan al-Ulama, n'a que 27 ans. Ils ont d'ailleurs aussi une Ministre de 30 ans en charge des voyages sur Mars qu'ils veulent atteindre avec une sonde inhabitée en 2021.

Arabie Saoudite

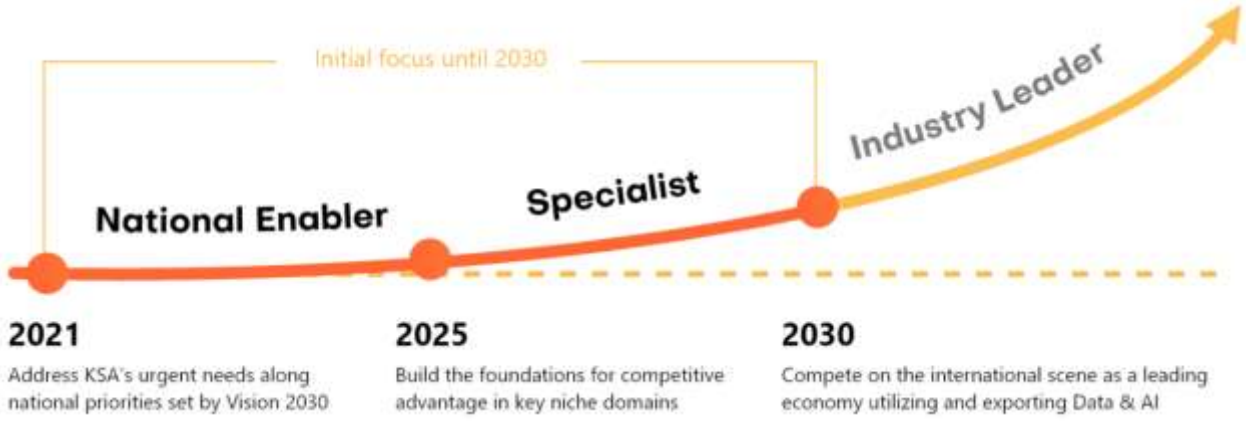
En novembre 2020, le pays annonçait son intention d'investir $20B dans l'IA et la data d'ici 2030. Le plan NSDAI (National Strategy for Data & AI) s'intègre dans le plan "Vision 2030" lancé par MBS, le turbulent prince héritier saoudien. Ils ont déjà une « Autorité des données et de l'intelligence artificielle » (SDAIA), créée en 2019. L'annonce présente l'IA comme une composante alternative au pétrole. Mais l'IA n'est pas une fin en soit. Ce qui compte est de créer des industries utilisant de l'IA, qu'il s'agisse d'éditeurs de logiciels, de services en cloud ou d'industries traditionnelles.

Le pays ambitionne en tout cas de créer 300 startups et de former 20 000 personnes dans l'IA et la data en 10 ans. Les $20B additionnent en fait de l'argent public et des investissements privés à venir. Les marchés visés sont l'éducation, le secteur public, la santé, l'énergie et la mobilité.

[2051] Infosys essaye de se positionner comme fournisseur de plateforme de développement d'application d'IA, avec le framework Ktri, codéveloppé avec Poyry et Nokia, deux sociétés finlandaises. Voir la brochure le décrivant. Ktri semble cibler les applications liées aux objets connectés dans l'industrie.

[2052] Voir UAE Strategy for Artificial Intelligence, octobre 2017 créé dans le cadre du plan UAE Centennial 2071.

Le tout s'accompagne d'un plan « open data » gouvernemental. Ils veulent entrer dans le club des 20 premiers pays du monde en IA en publications dans des revues à comités d'auteurs[2053].



Iran

De l'autre côté du golfe persique se trouve l'**Iran**. Sa population est l'une des plus éduquées dans l'enseignement supérieur du Moyen-Orient. C'est le 16ième pays du monde en termes de publications scientifiques dans l'IA.

Mais l'isolement économique de l'Iran, tout du moins vis-à-vis des pays occidentaux, s'est aggravé avec la sortie des USA des accords du JCPA (Joint Comprehensive Plan of Action) en 2018. Cela impacte par ricochet les partenariats économiques entre l'Europe et ce pays. Résultat, le pays ne peut se tourner que vers l'Asie, qui n'est pas un marché évident à aborder pour eux côté logiciels et Internet, du fait d'acteurs locaux bien implantés comme c'est le cas en Chine.

Le monde musulman n'est pas plus facile d'abord car l'Iran représente la minorité chiite face à la dominance du sunnisme. Qui plus est, les exportations du pays dépendent trop du pétrole et du gaz, à 79%.

[2053] Voir Saudi national AI strategy sets $20 billion investment target, octobre 2020 ainsi que Voir Realizing our best tomorrow strategy initiative, octobre 2020 (33 pages) qui démarre, tradition oblige, par deux portraits pleine page du roi et du prince-héritier..

IA et entreprise

Les entreprises de toutes tailles sont sous le feu de l'incantation de l'inévitabilité de l'IA. Une fois qu'elles ont décidé de faire quelque chose, reste à déterminer quoi, pourquoi, dans quel ordre, comment, avec qui et pour obtenir quels résultats.

En effet, le potentiel de l'IA est infini : on a l'embarras du choix des projets à mener. Il faut gérer les priorités. L'un des critères de sélection des projets est la disponibilité des données nécessaires à l'entraînement des IA. Contrairement à de nombreux projets numériques, l'adoption de l'IA passe encore plus par des tâtonnements et de l'expérimentation. Les compétences en IA étant rares, les entreprises vont se tourner naturellement vers des spécialistes, de grandes entreprises, des startups et/ou des prestataires de services qu'il faudra choisir en se créant de nouveaux repères et méthodes de sélection[2054].

Discours

En cinq ans, l'IA est passée du statut de paria de l'informatique à la tendance numéro un du numérique, alimentée par les performances médiatiques des GAFAMI et par l'actualité des startups. L'IA est devenue leur buzzword principal même s'il est maintenant concurrencé par celui de la Blockchain. D'ici peu, l'IA sera d'une banalité affligeante car largement répandue dans l'IT.

L'effet de suivisme était patent chez tous les cabinets de conseils et d'analystes qui ont tous publié leurs livres blancs sur l'IA[2055] entre 2015 et 2018. Nombre d'entre eux sont lénifiants, rappelant les définitions de l'IA (machine learning, deep learning, vision, langage…) et présentant quelques études de cas pas toujours bien documentées. Les chatbots ont souvent la part belle dans l'histoire car ils rentrent dans la vulgate de la transformation digitale et de la relation client. Selon le **Gartner**, l'IA était la première des trois grosses tendances de technologies émergentes en 2017 et en 2018[2056], avec la réalité mixte et les « plateformes digitales », c'est-à-dire le reste, avec dans le même sac, la 5G (2020…), les plateformes d'objets connectés, la Blockchain et les ordinateurs quantiques, dont la maturité n'est pourtant pas du tout au niveau de celle de l'IA.

[2054] Voir ce guide à l'attention des DSI du Gartner Group, A CIO's Guide to AI, 2018 (13 pages).

[2055] Par exemple, le livre blanc proposé dans Intelligence Artificielle : ce qu'il faut savoir pour l'expliquer à son responsable, de MagIT, 2018 (25 pages) est en fait un publi-rédactionnel pour valoriser une étude de cas associant Dataiku et PriceMoo !

[2056] Voir Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017, août 2017. C'était toujours le cas en 2018, voir : 5 Trends Emerge in the Gartner Hype Cycle for Emerging Technologies, 2018.

En 2018, l'IA était considérée comme une technologie en pleine phase de démocratisation, même si dans la réalité, cette démocratisation prendra encore quelques années. 2019 était l'année des débuts d'une adoption généralisée, en pièce détachée. 2020 aura peut-être été l'année des retours d'expérience, de la rationalisation et des déploiements opérationnels[2057].

D'autres sont d'avis qu'il ne faut pas se presser et prendre donc son temps, voir ne pas se lancer bille en tête dans un déploiement tout azimut de l'IA dans l'entreprise. Cela rappelle que l'adoption de toute nouvelle technologie n'est pas une fin en soi[2058].

Quelques analystes faisaient il y a peu de temps des prévisions de chiffre d'affaire pour le secteur de l'IA. Certaines se focalisent sur les industries numériques. D'autres, comme sur les objets connectés, portent sur le chiffre d'affaire de l'ensemble des industriels intégrant de l'IA dans leur offre. **PwC** prévoyait ainsi que l'IA devait faire croître le PIB mondial de $15,7T (trillions = mille milliards en américain) entre 2016 et 2030, soit 14% du PIB mondial actuel, ce qui est probablement exagéré[2059]. On nous avait fait le même coup avec les objets connectés ![2060]. A chaque fois qu'une évolution radicale du numérique arrive, elle va donc ajouter 15 points de PIB à l'économie mondiale ! En fait, elle transforme le PIB existant et l'industrie « primaire » du numérique est bien plus faible que cela. Les loisirs numériques représentent par exemple environ $1T, assez stables dans le temps[2061]. On flaire un peu de double booking dans ces prévisions mirobolantes ! Le logiciel influence d'ailleurs déjà une bonne partie de l'économie mondiale à lui tout seul.

Le propre des cabinets de conseil et analystes est d'instiller un sentiment d'urgence face aux révolutions technologiques. Ainsi, **Avanade** conseillait dans le rapport **Technology Vision 2017**[2062] « *aux entreprises d'intégrer dès maintenant l'intelligence artificielle (IA) pour rester compétitives* ». En précisant que « *les entreprises disposent d'une petite fenêtre de tir pour expérimenter et se familiariser avec les stratégies et les technologies qui préparent à l'arrivée de l'IA dans les pays industrialisés* ». Tout en recommandant aux entreprises « *d'agir avec responsabilité et d'adopter une éthique numérique* ». S'ensuivaient des recommandations qui correspondent probablement aux projets que l'ESN peut mener comme créer des applications avec des interfaces utilisateurs naturelles (vocales, tactiles, VR), des équipes de travail augmentées par l'IA et d'adopter ou de créer des plateformes.

Les discours autour du numérique peuvent être évidemment aussi anxiogènes. C'est le cas de cette étude de **Ricoh Europe** qui agitait l'épouvantail de la disparition pour les PME qui n'innoveraient pas, si possible avec du numérique et un zeste d'IA[2063]. Bref, tous les analystes s'accordent pour dire « *il faut y aller* ». Mais où, avec qui, comment et pour combien, c'est une autre histoire !

[2057] Voir Les projets autour de l'intelligence artificielle sont plus matures en 2020 par Melanie Wolkoff Wachsman, avril 2020, 10 Ways AI Can Improve Digital Transformation's Success Rate par Louis Columbus, avril 2020 et Leap And Learn: The Common Thread Of Artificial Intelligence Success Stories par Joe McKendrick, octobre 2020.

[2058] Voir Don't Rush Into AI Strategy Too Quickly, Expert Advises par Joe McKendrick, décembre 2019 et Non à l'introduction aveugle de l'IA dans l'entreprise par Gaspard Koenig, novembre 2019.

[2059] Voir AI to drive GDP gains of $15.7 trillion with productivity, personalisation improvements, juin 2017.

[2060] En 2015, General Electric prévoyait que les objets connectés généreraient $15T de croissance en 20 ans. **Cisco** évaluait cette croissance à $14 ,4T en moins de 10 ans. Voir The Internet of Everything is the New Economy, septembre 2015.

[2061] Ce $1T correspond au matériel, auquel il faut ajouter les télécommunications et les contenus, ce qui représente au moins $1,6T.

[2062] Voir Get ready for the AI first world, 2017.

[2063] Voir Sans innovation, un tiers des PME disparaîtra d'ici 2020 d'Ando Randrianarisoa, octobre 2018 dont le titre est évidemment biaisé. Il s'agit d'une enquête terrain d'un fournisseur, Ricoh Europe. Aucun lien sur l'étude n'est fourni dans l'article. Elle est en fait évoquée ici: Failure to innovate threatens to put a third of European SMBs out of business by 2020, octobre 2018, et téléchargeable sur ce lien. En subistance, 34% des patrons de 3300 PME européennes sondées dans 23 pays indiquaient que leur PME pourrait fermer s'ils n'innovaient pas assez vite. Cela ne veut pas dire que cela sera le cas ! Cette étude revient à dire que les PME dont les patrons sont incompétents ont plus de chances de mourir que celles qui ont un dirigeant à la hauteur. En guise de recommandations, l'étude propose aux PME d'améliorer leur relation client, d'utiliser des outils modernes en interne, d'automatiser les processus et de développer la créativité. Evidemment, les solutions de Ricoh sont là pour y pourvoir !

Dans « 7 AI myths », Robin Bordoli de la startup **CrowdFlower**[2064] (vidéo), synthétise bien les lieux communs sur l'IA que les entreprises doivent comprendre et éviter[2065] :

- **L'AI est magique et le deep learning peut résoudre tous les problèmes**. Non. L'IA, ce sont des données d'entraînement, des mathématiques, des probabilités et beaucoup d'itération avec de l'intervention humaine. Il n'existe pas de génération spontanée d'IA !
- **L'AI est réservée à une élite technologique** et aux GAFA. Elle est exploitable par toutes les entreprises, notamment via les nombreuses ressources disponibles dans le cloud. Mais il est vrai que l'outillage de l'IA nécessite encore des compétences de bas niveau qui ne sont pas encore suffisamment démocratisées. Les outils d'accès à l'IA montent cependant progressivement en niveau d'abstraction pour réduire cette complexité.
- **L'IA est dédiée à la résolution de gros problèmes** valant des milliards d'Euros. Ce document montre qu'il n'en est rien et que les entreprises de tous les secteurs d'activité sont concernées. Dès qu'il y a de la donnée, du langage, de l'image et des règles métier, l'IA peut faire quelque chose.
- **Les algorithmes sont plus importants que les données**. L'expérience montre que les deux sont aussi importants l'un que l'autre, sans compter le rôle des processeurs. La performance des algorithmes joue un rôle clé dans la qualité des résultats dans le deep learning, et surtout dans la rapidité de la phase d'entraînement des modèles. Mais il faut disposer de données de qualité et de processus pour les collecter et les préparer.
- **Les machines sont meilleures que les Hommes**. Non, car les machines ont presque toujours besoin d'interventions et d'informations d'origine humaine. Leur intelligence est alimentée par l'expérience et l'intelligence humaines comme dans les images de radiologie taggées par des radiologues ou des cancérologues. De plus, des IA couplées à des humains sont supérieures aux IA ou des humains seuls. Enfin, les hommes et les machines n'ont pas les mêmes capacités et se complètent.
- **Les machines vont remplacer les Hommes**. En fait, les machines augmentent les capacités humaines et réciproquement. Les métiers qui vont être entièrement remplacés par des IA ou des robots sont assez rares. Et lorsqu'il a remplacement, cela signifie souvent que l'on a déplacé du travail humain ailleurs, comme chez ces petites mains qui font de la labellisation de données où les clients qui prennent en charge une partie du processus métier.
- **L'IA, c'est du machine learning ou du deep learning**. Non. Il existe plein de techniques pour faire de l'IA, notamment autour de l'IA symbolique et du raisonnement automatisé, voire des réseaux multiagents et des outils de simulation. L'actualité les a mis de côté à cause du raffut autour du deep learning. Mais celui-ci a des limites. Les meilleures solutions d'IA intègrent et assemblent plusieurs techniques différentes.

En 2021, l'adoption de l'IA sortait d'une phase de maturation : les projets lancés par les entreprises sont nombreux et variés. Leurs résultats le sont tout autant. Les retours d'expérience rappellent que l'IA miracle n'existe pas et qu'elle est le fruit d'un travail humain significatif, de l'ingénierie des logiciels à la préparation des données[2066]. On découvre que l'IA n'est pas une fin en soi mais une boîte à outils qui complète les outils existants.

[2064] CrowdFlower (2007, USA, $58M) est une startup qui propose des outils d'exploitation des données pour alimenter des solutions de machine learning. Il automatise les business process amont et aval et met les utilisateurs dans la boucle pour affiner les données et les modèles.

[2065] Voir aussi le plus récent Gartner démystifie cinq idées fausses sur l'intelligence artificielle, par Patrick Ruiz, février 2019.

[2066] Voir Pourquoi se précipiter dans l'IA est un mauvais calcul, par Bob Violino, décembre 2018, qui évoque la question de la qualité des données qui alimentent le machine learning et L'adoption de l'IA et du machine learning prend du retard, par Tom Foremski, avril 2019.

Rares sont les solutions logicielles 100% IA. La plupart intègrent des briques traditionnelles qui sont couplées à des briques d'IA. L'utilisateur ne fait d'ailleurs pas trop la différence sauf lorsque l'IA implique du traitement de la parole ou des images.

Les entreprises n'ont pas encore toutes adopté une « stratégie IA »[2067]. Et pour cause, elles n'ont pas forcément de véritable stratégie numérique. L'IA n'est que l'une des composantes technologiques d'une stratégie numérique.

IBM présentait en janvier 2020 les résultats de son étude mondiale From Roadblock to Scale : The Global Sprint Towards AI qui porte sur d'adoption de l'IA dans les entreprises et son volet français résultant du sondage de 501 dirigeants d'entreprises. Que découvre-t-on ? Que deux tiers des entreprises explorent (36%) ou mettent déjà en œuvre (30%) de l'IA. Les obstacles sont le manque de compétences (27%), le cloisonnement des données (26%) et le manque d'outils (22%). Et deux tiers des répondants insistent sur l'impératif de la confiance dans les solutions développées.

McKinsey publiait de son côté en novembre 2020 son enquête « State of AI » dans laquelle les clients interrogés indiquaient qu'en majorité, les projets d'IA lancés étaient bénéfiques pour leurs business[2068]. Dans le même temps, un rapport du **BCG** et de la **MIT Sloan Management Review** indiquait plus ou moins le contraire, 10% des répondants ayant identifié un impact business significatif du déploiement d'IA. Leur recette ? L'intégration de l'IA dans une stratégie de performance globale. Surprenant n'est-ce pas [2069]?

Elle va se fondre progressivement dans les briques numériques de l'entreprise aux côtés des objets connectés, de la 5G et de la Blockchain pour ne prendre que quelques exemples d'actualité[2070]. Et si l'IA est déployée avec discernement, elle pourra même améliorer les conditions de travail et l'intérêt de nombreux métiers existants[2071].

Méthodes

Les dirigeants doivent aller au-delà du « *j'ai entendu parler de l'IA mais je ne sais pas quoi faire pour l'adopter* »[2072].

Dans l'IA comme dans de nombreuses nouveaux vagues technologiques, l'innovation va passer par le croisement d'une analyse de besoins mal traités, des attentes des clients, des inefficiences connues de l'organisation, et des potentialités technologiques à portée de main.

Il faut donc avoir comme bagage de départ une certaine compréhension de ce que les différentes briques et outils de l'IA pourraient apporter à l'entreprise. Il faut aussi en connaître les contraintes actuelles, techniques et économiques.

[2067] Voir Intelligence artificielle : une priorité, mais peu de stratégies d'entreprise par Christophe AUfray, juillet 2019.

[2068] Voir The state of AI in 2020, novembre 2020. Cela faisait suite à une enquête plus mitigée de 2019 : Global AI Survey: AI proves its worth, but few scale impact par McKinsey, novembre 2019.

[2069] Voir Are You Making the Most of Your Relationship with AI?, octobre 2020.

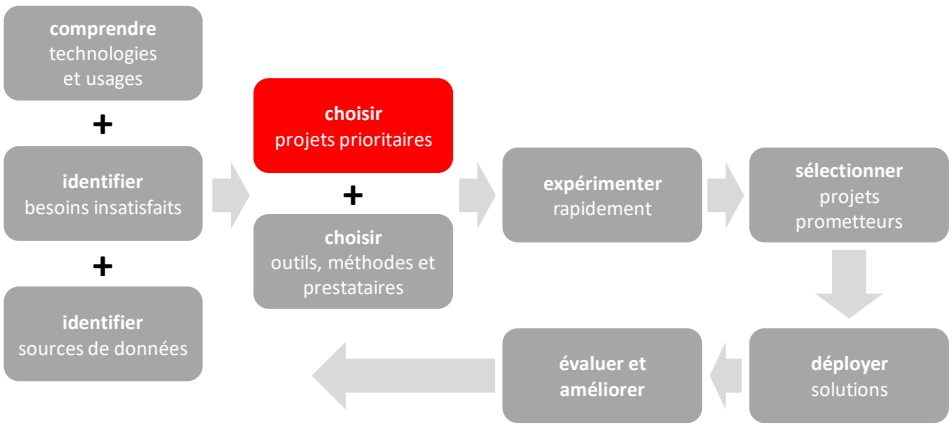
[2070] Voir The disappearance of AI par Kurt Cagle, janvier 2019. Et aussi Does AI Really Matter? par Kathleen Walch, mars 2020 qui fait écho au fameux « IT doesn't matter » de Nicholas Carr de 2003. Ici, la réponse est oui, l'IA a de l'importance. Mais, dira-t-on, au même titre que les grandes révolutions numériques de ces dernières décennies.

[2071] C'est la thèse de The AI Paradox: Four Benefits We Didn't See Coming par David Karandish, mars 2020. L'auteur évoque l'amélioration de la créativité et des conditions de travail grâce à la suppression de tâches fastidieuses, l'amélioration de la diversité si les systèmes de recrutement exploitent de l'IA pour débiaiser les recrutements humains, et la création de nouveau rôles.

[2072] Voir 4 tips to stop reading about AI and start doing AI - How to use AI as a tool in your business de Jana Eggers, janvier 2018.

Il faut ensuite évaluer les données disponibles qui pourraient alimenter des solutions d'IA.

Leur volume, leur origine et leur qualité jouent un rôle important dans la qualité d'une solution d'IA bâtie avec[2073].



Le tout peut être touillé dans des réunions d'idéation, sur paper boards et autres Post-it avec les parties prenantes. Il vaut mieux avoir d'abord mis les participants à niveau sur ce que l'IA permet de faire, par exemple avec une séquence d'acculturation avec un expert IA à la séance de brainstorming, afin de délimiter ce qui peut être pris en charge par l'IA et ce qui relève du mythe.

La stratégie « données » de l'entreprise sera affectée par l'IA. Les applications de l'IA impacteront la stratégie et les méthodes d'acquisition de données via l'IOT ou d'autres moyens et même sur la stratégie d'open data de l'organisation.

Enfin, on fera le tri des projets potentiels pour choisir ceux qui sont les plus pertinents en fonction de grilles de choix classiques (quick win, coût modéré, avantage concurrentiel…) puis on passera en phase d'expérimentation.

Plus les équipes seront au fait du potentiel de l'IA, plus le nombre de projets potentiels à lancer sera important. On peut passer très rapidement du « que faire avec l'IA » avec un trop plein d'idées. Il faut alors en revenir aux choix business classiques et retenant les projets qui ont le meilleur impact sur les priorités de l'entreprise[2074].

La notion de « preuve de concept » (PoC) est particulièrement valable dans l'IA même si on ne passe pas par une startup pour la mener. La raison est que la majeure partie des solutions d'IA ne génèrent pas un résultat déterministe. Il faut les expérimenter pour en évaluer la qualité. Les solutions d'IA génèrent un taux d'erreur qu'il faut faire descendre aussi bas que la technique et l'équation économique le permettent, et dans la mesure du possible, en-dessous du taux d'erreurs humaines habituelles.

A la suite des expérimentations, il y aura du déchet. Sinon, il n'y aurait pas de processus d'innovation à proprement parler. Seuls les PoC réussis donneront lieu à un déploiement. Et il faudra y consacrer des investissements[2075].

On rebouclera alors la boucle pour améliorer les projets déployés et découvrir la potentialité de nouvelles technologies d'IA apparues depuis le début du cycle[2076].

[2073] Cet impératif sur les données est rappelé à qui mieux mieux dans tous les discours. Voir par exemple How AI changes the rules par SAS, 2020 (24 pages), qui recommande d'impliquer les CEO dans l'adoption de l'IA et de s'intéresser aux données.

[2074] Voir The New Age Enterprise - Enabled by AI par Sameer Dhanrajani, janvier 2019 évoque en le survolant l'impact stratégique de l'IA dans les entreprises : dans la production, dans la logistique et la supply-chain, dans le marketing et les ventes, dans l'offre elle-même, dans la RH.

[2075] Voir IA: rajoutez un zéro! de Sylvain Duranton, BCG, octobre 2018 qui insiste sur l'importance des déploiements après les projets pilotes. C'est là que beaucoup d'énergie doit être investie.

[2076] J'ai retrouvé a posteriori des traces de cette méthode dans Artificial Intelligence for the Real World de Thomas Davenport et Rajeev Ronanki dans HBR, janvier 2018.

Les cycles de développement vont aussi évoluer. Dans l'IA, la mise au point des modèles d'entraînement de machine learning et de deep learning génère des allers et retours plus long que la correction de bugs classiques.

L'entraînement d'un modèle peut être très long, même avec les batteries de serveurs les plus puissantes. On ne débogue pas de tels modèles de manière aussi interactive que les langages interprétés du web et même que ceux qui sont compilés.

D'un autre côté, l'expression de ces modèles avec les langages de programmation courants tels que Pythoon et R, couplés à des SDK comme TensorFlow est plus concise. Il n'y a rien d'automatique dans tout cela malgré les lieux communs sur l'IA.

La mise en œuvre de solutions de machine learning requiert d'expérimenter divers modèles de représentation des données, de segmentation, de prévision. Celui du deep learning passe par la définition de modèles en couches empilés dont la forme et le dimensionnement dépend des données à analyser : images, voix et textes[2077].

D'autres recommandations vont dans sens de l'action, comme cette méthode en six étapes qui comprend : la compréhension des bénéfices de l'IA, l'analyse des meilleurs pratiques du marché, l'assemblage de compétences internes et externes, l'approche transversale dans l'entreprise, l'accompagnement externe et le démarrage expérimental avant l'industrialisation[2078].

En 2016, le CIGREF prodiguait quelques recommandations sur la méthode à employer pour adopter l'IA dans son organisation, qui sont toujours d'actualité et que je vais commenter[2079] :

- **Affecter un budget dédié à l'IA** : cela peut avoir du sens pour mettre en place les outils génériques utilisés par les premiers projets. C'est donc une optique de mutualisation a priori. Faut-il des budgets pour les projets eux-mêmes ? Je ne le pense pas. C'est le business qui décide des priorités et l'IA est une technique parmi d'autres techniques avec l'IOT, la BlockChain, la mobilité ou le cloud pour réaliser ces projets.
- **Passer à l'internet 4.0** : IA, algorithmes prédictifs : ce sont des termes à la mode. Les algorithmes prédictifs font partie des différentes techniques utilisables mais ne sont pas les seules.
- **Engager un roboticien** dans des équipes IT pour passer en 4.0 : si on est dans des métiers « physiques ». Pour la banque, cela a peu de sens, à moins que cela s'applique aux notions de Robotic Process Automation qui sont liées à l'automatisation de processus métiers de cols blancs et de backoffice.
- **Développer des systèmes de Machine Learning** : il faut s'approprier les outils du machine learning et du deep learning pour en tirer le meilleur parti selon les besoins dans des projets. Si on peut s'affranchir de la complexité de ces outils pour bâtir ses solutions, autant en profiter. De plus en plus d'outils vont dans ce sens pour alléger la charge mathématique et algorithmique des concepteurs de solutions de machine learning. Cela correspond à la vague des outils AutoML, qui aident les créateurs de solution à choisir automatiquement le meilleur modèle de machine learning en fonction du jeu de données dont on dispose.
- **Suivre les tutoriels de TensorFlow** : OK car c'est l'outil générique le plus utilisé pour créer des solutions de machine learning et de deep learning. Et il fonctionne en embarqué ou sur serveurs, on premise ou dans le cloud. Mais ce n'est pas le seul. Il y a aussi des outils plus flexibles comme PyTorch et Caffe2. Et on peut s'approprier le machine learning via des outils d'analyse de données adaptés qui ne nécessitent pas forcément de faire de la programmation.

[2077] Voir cet inventaire des changements qui affectent le développement logiciel dans les entreprises : How AI Will Change Software Development And Applications, par Diego Lo Giudice de Forrester, octobre 2016.

[2078] Voir Getting Intelligent About Artificial Intelligence:- 6 Ways Executives Can Start de Davia Temin, décembre 2017.

[2079] Voir Gouvernance de l'intelligence artificielle dans les grandes entreprises, Cigref, septembre 2016.

- **Développer la culture des APIs en interne** : oui, et indépendamment de l'IA, histoire de bien décomposer le système d'information en services interopérables, de favoriser la publication et l'usage d'open data, de transformer son activité en plateforme ouverte aux autres entreprises. C'est aussi du « business design ».
- **Sensibiliser les Métiers et Fonctions aux enjeux de l'IA** : en effet, et en leur faisant des propositions, en croisant les capacités d'usage et les besoins des métiers. L'évangélisation interne des métiers aux capacités de l'IA est une phase importante de l'acculturation. Cela donne des idées. Cette acculturation doit aussi gommer une bonne part de la dynamique anxyogène qui entoure l'IA. Elle doit aussi gommer l'impression que l'IA est magique.
- **Développer une communauté autour de l'IA et échanger** : idem, comme sur tous les sujets technologiques du moment, l'IA étant d'ailleurs souvent un outil associé à ces différents domaines.
- **Supprimer les « points de douleur » dans l'entreprise** : ce n'est pas spécifique à l'IA. C'est une approche d'innovation passant par l'identification de problèmes à résoudre.
- **Créer des boîtes noires logiques qui gardent en mémoire l'IA** et avoir la possibilité de la détruire (d'effacer l'ensemble des parcs) dans un souci de droit à l'oubli. Effacer les données d'un SI n'est jamais un véritable problème. Les conserver en est un ! La mémoire des IA, notamment à base de deep learning, est située dans les données d'entraînement et dans les paramètres des réseaux de neurones entraînés. Il est important de bien conserver les jeux de données d'entraînement, ne serait-ce que pour pouvoir auditer les systèmes qui deviendraient défaillants.

En octobre 2017, le Cercle de l'IA du CIGREF complétait ce premier rapport avec un nouveau document qui donne la part belle aux leçons tirées de l'expérience de la mise en œuvre de chatbots clients, avec deux exemples : ceux d'Orange et le cas d'usage dans la RH. Il fait de nombreuses recommandations sur leur mise en œuvre, autour des questions pratiques des données qui l'alimentent et de l'éthique[2080].

Le document se fait aussi l'avocat d'une IA ouverte, permettant aux entreprises de faire dialoguer leurs IA entre elles, dans une approche voisine de celle de l'open data. L'architecture proposée est voisine de celle de l'open data accessible via des APIs avec des IA gérant leurs propres îlots de données, qui sont exposées via des services d'accès. Reste à déterminer au cas par cas les données transmises entre IA, dans le respect de la vie privée des utilisateurs et des atouts stratégiques des entreprises. La manière dont l'architecture des SI à base d'IAs semble assez traditionnelle dans l'approche. Le principe même d'un « agent » dans l'IA est d'être relativement indépendant et d'évoluer en fonction de son environnement.

Le Rapport du CIGREF évoque aussi la question clé de la collaboration entre les entreprises et la recherche, en la confondant quelque peu avec les startups dont l'objectif est plutôt de créer des produits qui exploitent de la recherche existante.

[2080] Voir Intelligence artificielle dans les grandes entreprises, enjeux de mise en œuvre opérationnelle, Cigref, octobre 2017 (36 pages).

La vitesse de transfert des travaux des chercheurs vers les applications d'IA est plus rapide qu'auparavant, ne serait-ce que parce que nombre de travaux de recherche s'appuient sur la publication de code exploitant des jeux de données standards (MNIST, ImageNet, WordNet), faciles à reproduire par d'autres développeurs.

En octobre 2018, le Cigref remettait le couvert en se focalisant sur la question de la gouvernance des données avec des témoignages d'Orange, Engie, Enedis, La Poste, PSA, Saint Gobain, Air France KLM, Hager, Caisse des Dépôts, de Pôle Emploi et de la Sacem. Le rapport propose aussi diverses formes d'organisation pour adopter l'IA dans l'entreprise (communautés d'experts, challenges d'IA, data innovation labs, etc)[2081].

Projets

Supposons qu'un besoin soit bien identifié, que les données soient disponibles et que les technologies de l'IA puissent apporter une solution. Une fois ceci qualifié, on peut rentrer en mode projet.

Va-t-on mener plusieurs projets pilotes en parallèle avec différents frameworks d'IA tels que ceux qui sont cités auparavant, et comparer ensuite les résultats ? Ce ne serait pas raisonnable et surtout, ce serait trop coûteux. Il vaut mieux se faire conseiller pour choisir les bons outils et ensuite mener son expérimentation.

Un projet d'IA d'entreprise a un petit côté "recherche applicative", qu'il s'agisse d'un chatbot, d'une application industrielle, d'un système de vision artificielle ou d'un outil d'analyse de données pour faire du prédictif.

Pour prendre ce dernier exemple, on ne va pas juste alimenter une bébête de machine learning ou de deep learning avec un tombereau de données et attendre un beau résultat à la sortie d'un tuyau. Il va falloir d'abord extraire et préparer les données, les nettoyer, les filtrer, savoir ne conserver que ce qui est pertinent.

On va ensuite paramétrer les outils de machine learning ou deep learning en fonction des algorithmes à utiliser. Comme nous l'avons vu dans ce document, il n'existe pas une technique unifiée de machine learning ou de deep learning, mais des dizaines de variantes ! Puis on va observer les résultats. Ils ne seront pas forcément probants du premier coup. Il faudra reboucler sur les données et le paramétrage pour affiner le modèle. Et il faudra aussi bien visualiser les résultats pour qu'ils soient compréhensibles. La partie « dataviz » d'une application d'IA à base de machine learning est aussi importante que les algorithmes retenus.

On appréciera alors la qualité des résultats. Les techniques de machine learning et de deep learning génèrent rarement des résultats exacts à 100%. Il y a toujours un taux d'erreur, que l'on minimise avec l'expérience et que l'on cherche à faire descendre en-dessous d'un niveau acceptable. Comme la variété des échanges typiques acceptables dans un chatbot, le taux d'erreurs d'un système de reconnaissance vocale, ou celui de l'identification de pathologies dans de l'imagerie médicale. A ce jour, les solutions les plus avancées dans ce dernier domaine génèrent un taux d'erreur plus faible que celui des spécialistes ! C'est donc acceptable !

Dans les cas où l'explicabilité du résultat est plus importante que sa fiabilité, comme dans certains métiers très régulés, on pourra être amenés à éviter les solutions à base de deep learning non explicables.

[2081] Voir L'intelligence artificielle en entreprise, Cigref, octobre 2018, (40 pages).

Un benchmark pourra éventuellement avoir lieu pour comparer un projet mené en mode “IA” et un projet mené avec des outils traditionnels de data mining. Si ceux-ci peuvent donner des résultats convenables sur des données chiffrées, ils ne sont maintenant pas du tout à la hauteur pour traiter des données images/vidéo/audio, là où le deep learning est devenu indispensable.

L’IA pose une question habituelle : faut-il développer une solution sur mesure où existe-t-elle sur étagère ? Les éditeurs de logiciels d’IA actuels peinent à créer des solutions véritablement génériques, sauf lorsqu’ils créent des outils de développement. Ils proposent souvent leur solution en mode projet ou service outillé. La raison est que leurs outils n’arrivent pas encore à s’adapter aux sources et structures de données utilisées par les clients, qui varient significativement d’un cas à l’autre. On peut prévoir que cela va s’améliorer à terme, témoignant d’une maturation des projets d’IA en entreprise. Mais on risque de rester assez longtemps comme dans la situation du déploiement d’un ERP qui nécessite beaucoup de développements personnalisé[2082].

L’un des points clés pour les entreprises est de passer du projet pilote (PoC) à la production au déploiement à grande échelle. Il y a pas mal de perte en ligne, mais c’est normal lorsque l’on est aux débuts d’un cycle d’innovation.

Ainsi, une étude mondiale aurait déterminé en 2019 que 50% des projets d’IA lancés par les 25% de grandes entreprises qui ont adopté l’IA ont périclité, le plus souvent faute de disposer des compétences adéquates et du fait d’attentes trop grandes[2083]. En effet, l’IA ne fait pas des miracles ! Si les données d’entraînement ne sont pas de qualité et quantité suffisante, les IA connexionnistes ne fonctionneront pas bien. Les applications d’IA requièrent souvent un gros travail d’intégration et aussi, du bidouillage ! Cela reste artisanal.

L’autre source potentielle d’échecs, moins connue, consiste à ne pas assez se soucier de l’utilisateur et/ou du client, avec des interfaces utilisateurs mal conçues ou tout simplement avec des automatismes qui n’apportent pas grand-chose[2084]. Les basiques d’un bon projet restent donc toujours valables : il faut impliquer les utilisateurs et faire du design ! Et donc, éviter de supprimer les humains de la boucle [2085] ! Au passage, en précisant bien aux utilisateurs qu’ils ont affaire à une IA lorsque c’est le cas[2086]. L’acceptation des nouveaux outils par les utilisateurs est toujours un gage de succès d’un projet.

Benchmarks

Le premier élément d’un benchmark consiste à analyser les études de cas du marché qui sont voisines des projets que l’on souhaite lancer. En matière d’IA, il faut être particulièrement vigilant. Nombre d’études de cas mises en avant par des fournisseurs de technologies exagèrent les résultats voire travestissent entièrement la réalité des projets.

Points de vigilance

Quelques points de vigilance sont à observer : qualifier et quantifier les données qui alimentent les systèmes ainsi que leur origine et leur fiabilité, décortiquer les outils logiques, logiciels et matériels utilisés, et analyser les résultats. Enfin, la structure de coût, sa maintenabilité et la durée du projet sont à intégrer dans l’évaluation.

[2082] Voir notamment Build, buy, or both? The AI implementation conundrum, de Pedro Alves, octobre 2018.

[2083] Voir This Week In AI Stats: Up To 50% Failure Rate In 25% Of Enterprises Deploying AI par Gil Press dans Forbes, juillet 2019.

[2084] Voir Artificial Intelligence (AI): What About The User Experience? par Tom Taulli dans Forbes, avril 2019.

[2085] Voir How to prevent embarrassment in AI par Cassie Kozyrkov, juillet 2019, sur les aléas des projets d’IA comme les chatbots qui ne fonctionnent pas bien, le machine learning entraîné avec des données insuffisantes et le besoin de conserver l’humain dans la boucle.

[2086] Voir Why AI systems should disclose that they’re not human Otherwise, they could mislead and deceive people par Alec C.Engler, janvier 2020.

Il faut aussi avoir une vision globale d'un projet. Par exemple, un chatbot marketing utilisé dans la relation client doit être évalué sur son impact global sur la satisfaction client et pas seulement sur son impact sur le coût du support commercial ou technique.

Dans le contexte d'un projet d'entreprise, un projet d'IA démarre souvent avec des données et si possible avec de gros volumes de données, même si la tendance de l'IA frugale est à l'utilisation de plus faibles volumes de données. Le volume et la qualité des données sont clés pour bien entraîner un moteur de deep learning. C'est l'une des raisons de la force des GAFA : ils ont naturellement accès à d'immenses volumes de données liées aux actions des utilisateurs de Google Search, Facebook, iOS, Android, SIRI, Amazon Alexa, etc. Les sociétés qui déploient de gros volumes d'objets connectés ont aussi accès à des données intéressantes à exploiter.

Un benchmark d'entreprise doit donc partir d'un ou de jeux de données dont on veut extraire quelque chose.

Il faut bien évidemment se poser la question de ce que l'on veut en faire. Au départ, on ne sait pas trop. L'entreprise dispose par exemple d'une base de données du comportement de ses clients et voudrait l'utiliser pour identifier les clients à potentiel d'upsell ou de cross-sell (ventes additionnelles), ou au contraire, ceux qui peuvent générer du churn (abandonner l'offre). Elle peut aussi vouloir déterminer les actions à mener pour optimiser un système complexe : client, production, autre.

L'IA peut aussi servir dans tout un tas de domaines : dans la robotique (qui intègre généralement tout un tas de briques technologiques : vision artificielle, mécanique, systèmes experts, etc), dans la relation client, pour créer des solutions de recommandation, pour analyser des tendances, pour analyser l'image de l'entreprise dans les médias et les réseaux sociaux. Etc. Et la gradation est forte entre générique et spécifique dans ces différentes solutions.

Des projets d'IA peuvent se passer de machine learning et de deep learning et s'appuyer sur des connaissances structurées et des moteurs de règles. C'est par exemple le cas pour créer des systèmes d'assistance à la maintenance industrielle. Dès lors que l'on manipule des données très structurées et une architecture de concepts, les outils de deep learning sont inadaptés. On se retrouve ici dans un domaine ancien, qui a connu ses heures de gloire pendant les années 1980, avec LISP et Prolog. Il n'est pas périmé pour autant, malgré tout le tintouin autour du deep learning, présenté à tors comme une sorte de solution universelle des besoins de l'IA. On va alors faire appel à des **BRMS**, des Business Rules Management Systems.

Conceptuellement, pour les entreprises qui disposent de gros volumes de données, l'IA constitue souvent un ensemble de techniques qui complète une longue lignée de technologies : les bases de données, la business intelligence, le big data, les data analytics et la data intelligence. C'est donc une évolution plus qu'une révolution pour elles.

Modèle d'étude de cas

Vu des clients, il est critique d'accéder à des études de cas de fournisseurs, histoire d'évaluer l'intérêt de lancer tel ou tel projet d'IA dans son entreprise. Voici une proposition de modèle de documentation d'étude de cas de projet intégrant de l'IA[2087]. C'est un modèle extensif qui sera probablement rarement complètement rempli. Peu d'entreprises ont envie de documenter leurs projets avec ce niveau de détails. Mais ces études de cas peuvent être réalisées par certains éditeurs pour des projets présentés "behind closed doors".

<u>Société cliente</u>

- Secteur d'activité.

[2087] Je reprends ici une proposition que j'avais publiée en décembre 2017 dans Modèle d'étude de cas de l'IA.

- Taille de l'entreprise. Bien préciser la taille de l'entité couverte par la solution. "Total" ou "Orange" n'est pas assez précis. On est souvent trompé par les études de cas qui ne précisent pas leur portée dans une très grande entreprise. Très souvent, les projets n'en concernent qu'une toute petite entité.
- Lieu, ce qui intéressant dans le cas de déploiements internationaux.

Solution

- Description métier du besoin et de la solution. Comment faisait-on avant ? Quelles techniques classiques étaient utilisées ? Quels étaient les surcoûts engendrés par l'existant ?
- Description technique de la solution. Quelles techniques d'IA intègre-t-elle : de l'IA symbolique (système expert, moteur de règle, logique floue), du machine learning, des réseaux de neurones simples, du deep learning, des réseaux convolutifs, des réseaux récurrents ou à mémoire, des techniques de traitement du langage.
- Copies d'écrans de la solution, vue de l'utilisateur. L'interface utilisateur d'une solution logicielle est aussi importante que sa fonction !
- Schémas fonctionnels, un diagramme des flux des données avec leurs sources étant indiqué.
- Périmètre de la solution : projet pilote ou déploiement industriel.

Données

- Nature, volume et origine des données d'entraînement puis de production. Quels capteurs les ont générées (logs Internet, objets connectés, …). Quelles données sont d'origine interne et externe à l'entreprise ? Quelles données exploitées relèvent de l'open data.
- Fréquence de la mise à jour opérationnelle des données. Comment le modèle est-il réentraîné avec l'arrivée de nouvelles données ?
- Taux d'erreur mesuré de la solution si applicable. Ce taux est mesuré après l'entraînement du système d'IA si celui-ci utilise du machine learning ou du deep learning.
- Anonymisation des données exploitées le cas échéant. Est-ce que les données qui alimentent le machine learning ou le deep learning sont bien anonymisées ou pseudonymisées (pour qu'un croisement de données ne permette pas l'identification d'utilisateurs, comme vu dans le thème de la *differential privacy*). Normalement, c'est toujours le cas. La solution met-elle en place des techniques de differential privacy ?

Fournisseurs

- Technologies. Au sens : logiciels de base (TensorFlow), d'infrastructure (Spark, Hadoop), progiciels divers et autres.
- Prestataires de services. En indiquant leur apport dans le projet.
- Ressources en cloud si pertinent. Et notamment, si des processeurs spécialisés (GPU ou neuromorphiques) sont utilisés, notamment pour l'entraînement d'un modèle de deep learning.

Dates

- Début du projet.
- Date des premiers tests opérationnels. Ce que l'on appelle un "PoC", pour proof of concept.
- Date de la mise en production. Et portée de la mise en production en nombre d'utilisateurs.

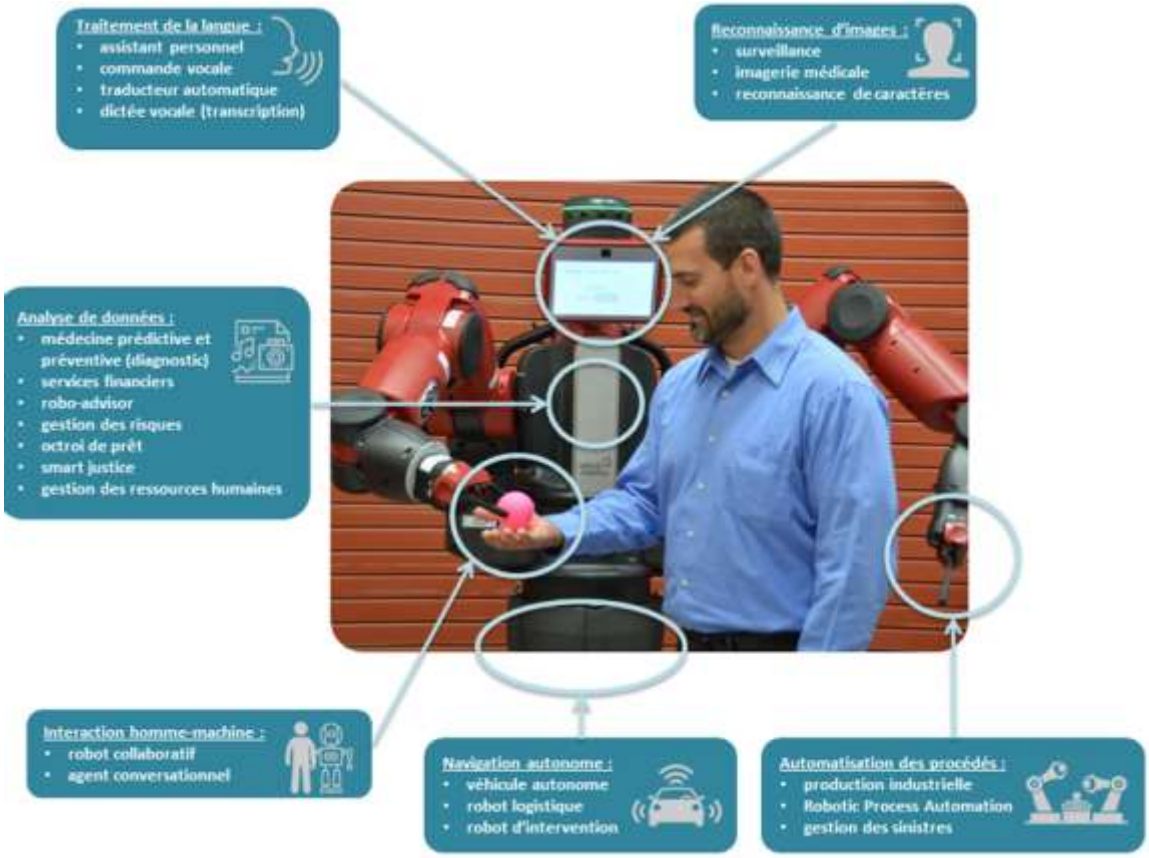
Economie

- Coût du projet. Ressources humaines consommées en interne et en externe pour créer la solution. Types de compétences : développeurs, data-scientists, etc.

- Durée d'entraînement des modèles, dans le cas de solutions à base de machine et de deep learning. Et avec combien de serveurs on-premise et dans le cloud et pour quel coût.
- Nombre d'utilisateurs de la solution, aujourd'hui et demain.
- Retour sur investissement. C'est la partie la plus difficile à mesurer sur de nombreux projets. Il faut pouvoir y intégrer l'ensemble des coûts relatifs au projet, y compris la formation des utilisateurs.
- Validation du projet au regard de la RGPD, la Règlementation Générale de la Protection des Données européenne qui est entré en vigueur le 25 mai 2018.

Bonne nouvelle, le **LNE** (Laboratoire National d'Essai) s'est lancé en 2018 dans les processus d'évaluation de solutions à base d'IA[2088]. Ils proposent cela sous forme de service aux entreprises et couvrent un large spectre de solutions d'IA (*ci-dessous*).

Cela comprend notamment les solutions de traitement de la parole et de reconnaissance d'images. Ils évaluent également le bon fonctionnement de robots, y compris de robots humanoïdes et de robots agricoles, ainsi que celui des capteurs utilisés dans les véhicules autonomes. Cette évaluation est aussi exploitable par les financeurs publics de l'innovation afin d'évaluer les résultats des projets financés de manière rationnelle et normée.



Du côté de l'outillage de benchmarks, on peut citer MLPerf, un outil open source qui permet de gérer ses benchmarks d'applications de machine learning[2089]. Il est exploitable pour les solutions de classification d'images, de détection d'objets, de reconnaissance de la parole, de traduction, de recommandation, d'analyse de sentiments et d'apprentissage par renforcement.

[2088] Voir Evaluer les Intelligences Artificielle, septembre 2018 et Évaluation des systèmes d'intelligence artificielle.

[2089] Voir A new benchmark suite for machine learning, mai 2018 et le site du projet : https://mlperf.org/.

Outils

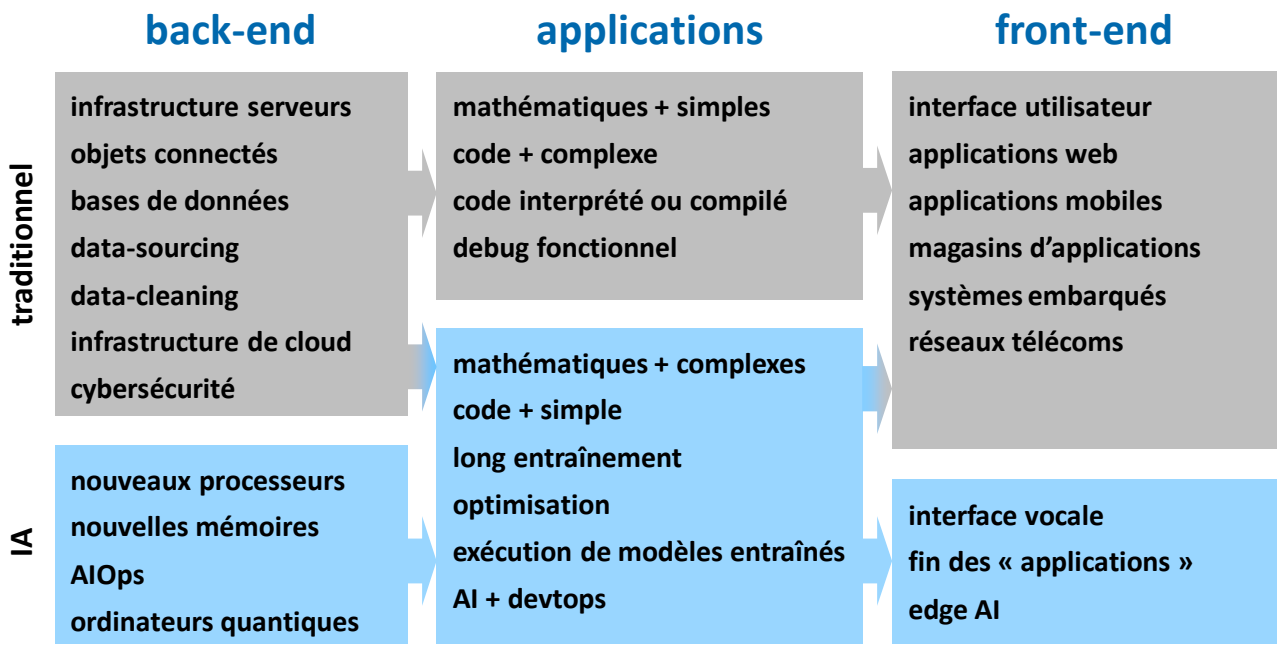
L'entreprise ou ses partenaires devront faire des choix d'outils pour mener leurs projets d'IA. Dans tous les domaines de l'IA, il y a déjà un énorme embarras du choix. La majorité des solutions logicielles de développement sont publiées en open source. Mais il y a des variantes dans leur mise en œuvre. Certains éditeurs proposent ainsi des briques propriétaires autour de leur souche open source.

Les acteurs du marché des outils logiciels de l'IA se rémunèrent avec du service, des solutions métiers payantes, des ressources en cloud, voir du matériel spécialisé.

Il va falloir déterminer où exécuter ses solutions. Si elles sont demandeuses de ressources machine importantes pour l'entraînement ou l'inférence de modèles, avec des serveurs à base de GPU, il sera censé de faire appel à des fournisseurs de telles ressources en cloud. Aujourd'hui, ils doivent être équipés de serveurs DGX-1 ou DGX-2 de Nvidia pour être à la page. Les fournisseurs doivent évidemment être à même de protéger les données de l'entreprise, même les jeux de tests. Ils fournissent des « cloud privés » adaptés à ce genre de besoins.

On fera aussi des choix de topologie d'IA, par exemple, en déterminant où sont réalisés les traitements. Dans certains cas, on les fera réaliser au niveau des capteurs, comme pour certains types de caméras de surveillance qui intègrent leur propres outils de détection d'intrusion et remontent des alertes via le réseau. Certains outils comme la bibliothèque TensorFlow sont conçus pour être exécutés indifféremment sur des objets ou sur des serveurs. Des architectures matérielles comme les GPU de Nvidia le sont tout autant avec des déclinaisons côté serveurs et pour l'embarqué (Jetson)[2090].

Néanmoins, les applications de l'IA ne vont pas révolutionner la totalité des systèmes d'information. Elles en exploiteront des briques existantes comme illustré dans le schéma ci-dessous. Ainsi, en amont des outils de machine learning et de deep learning se trouvent des bases de données et des serveurs d'infrastructure traditionnels.



[2090] Voir Architecting For AI, de Ann Steffora Mutschler, juillet 2018 qui relate un débat de spécialistes sur le sujet.

En aval, les applications proposées aux utilisateurs s'appuient toujours sur des interfaces, en général graphiques, adaptées aux micro-ordinateurs, mobiles ou systèmes industriels. La conception d'interfaces utilisateurs doit intégrer le besoin d'explications dans certains cas permettant de bien interpréter les résultats affichés[2091].

L'un des impacts de l'IA est de développer l'usage de la commande vocale. Il fait même émerger la notion de « VUI » pour Vocal User Interface qui décrit les techniques et bonnes pratiques de gestion de l'interaction vocale avec les utilisateurs. Au passage, les applications vocales intégrées dans les plateformes telles qu'Amazon Alexa sont utilisées de manière transparente par les utilisateurs, faisant disparaître la notion même d'application.

Dans quels cas fera-t-on appel à des chercheurs ? La question se posera lorsque l'état de l'art de l'IA sera insuffisant pour résoudre un problème donné. L'entreprise aura surtout besoin de « recherche appliquée » qui pourra être sous-traitée à un laboratoire de recherche (Inria, un IRT, un 3IA) ou à une équipe de post-docs in situ dans l'entreprise (via contrat CIFRE par exemple). Le point clé à gérer sera l'horizon de temps de ces chercheurs. Il devra être compatible avec celui du projet ! Mais dans la plupart des cas, on n'aura pas besoin de chercheurs, mais plutôt d'une ribambelle de compétences diverses pour gérer son projet (compétences métiers, data scientist, développeur, designer)[2092].

Financement

Il existe quelques moyens de bénéficier d'aides et prêts pour financer ses projets d'IA. Pour augmenter la compétitivité des entreprises et en particulier des PME et de ETI, L'Etat met à disposition des solutions de financement, de crédits d'impôts souvent complétée par un effort des collectivités notamment de la région.

La loi de finances 2019 permet aux PME de sur-amortir leur investissement à hauteur de 140% de leur valeur dans la robotique et dans la transformation numérique. Bpifrance finance à 50% les diagnostics de PME et ETI sur les opportunités que peut apporter l'Intelligence Artificielle.

La région Île de France propose avec le Pack IA d'accompagner de manière personnalisée des PME et ETI dans leur accès aux technologies d'IA qui va jusqu'à la mise en production de la solution.

Il est aussi possible de bénéficier du crédit impôt recherche y compris pour des dépenses de recherche externalisées, et pour les dépenses liées aux prototypes et installations pilotes de produits nouveau, le Crédit d'impôt innovation.

Compétences

Comme pour toute nouvelle technologie, les entreprises font généralement appel à des spécialistes de l'IA qui connaissent la diversité de ses nombreuses techniques et méthodes.

D'après le plan France IA du gouvernement publié en mars 2017, les principaux métiers (et compétence) qui devraient apparaître ou se généraliser autour de l'IA seraient :

- **Architecte en conception d'IA** : une fonction dans la lignée des architectes de systèmes d'information, qui requiert une vue globale des techniques d'IA et une capacité à les composer, autant dans les architectures logicielles, matérielles que cloud.

[2091] Voir à ce sujet Council Post: AI Is Outperforming Humans, So Why Don't Customers Trust It? par Abinash Tripathy, janvier 2020.

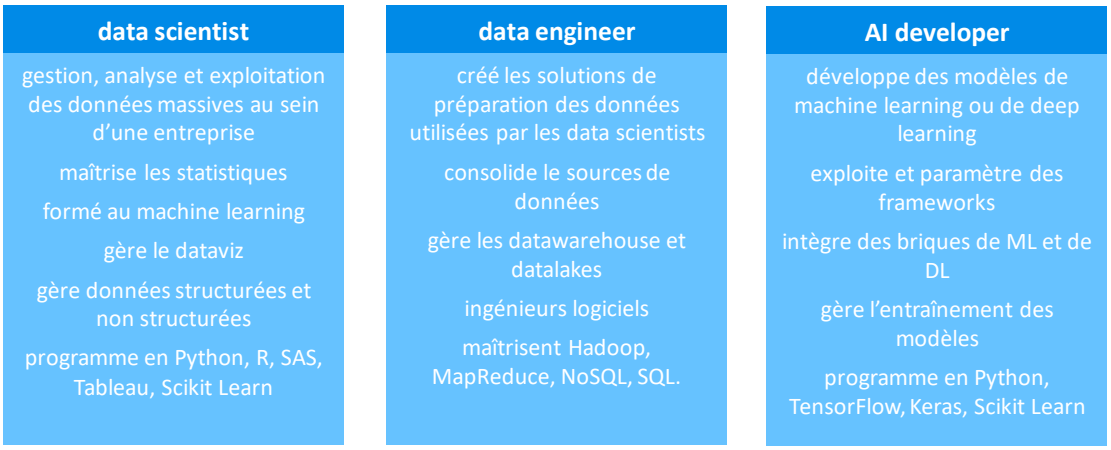
[2092] Dans Why businesses fail at machine learning, juin 2018, Cassie Kozyrkov de Google explique pourquoi vous n'avez pas forcément besoin de PhD pour votre projet d'IA mais d'un alignement de compétences, surtout métier.

- **Intégrateurs d'IA** : il s'agit de développeurs ayant une bonne compréhension et pratique des techniques de machine et de deep learning qui adapteront ces briques technologiques aux usages métiers. Ces développeurs doivent mieux maîtriser les mathématiques des modèles de machine et de deep learning. Ils vont faire plus de maths et moins de code [2093]!
- **Spécialistes métier** : que l'on retrouve habituellement dans les fonctions de MOA (Maîtrise d'ouvrage), qui ont une compréhension d'un métier et des données associées, et font le lien entre le besoin métier et les équipes techniques existantes, aident à sélectionner et utiliser les nouveaux outils embarquant une IA.
- **Concepteurs d'interactions avec les IA et robots** : qui maîtrisent l'utilisation de données comportementale et l'ergonomie pour concevoir et spécialiser les interfaces avec les utilisateurs moins qualifiés et les clients. Ils ont des compétences en design !
- **Entraîneurs d'IA** : moins ou pas qualifiés sur les techniques d'IA mais ayant une haute spécialité de leur métier et qui alimenteront en données de référence divers systèmes d'IA. C'est le cas par exemple des systèmes de traitement du langage ou de chatbots qui ont besoin de données textuelles types pour fonctionner.

Dans les petites structures telles que les startups, l'ensemble de ces activités sera concentrée sur un nombre réduit de spécialistes, et même parfois, un seul.

Le marché s'attend également à une forte demande de chefs de projets intervenant de manière transversale sur le développement, l'intégration, et la maintenance des systèmes d'IA, notamment dans les domaines du machine learning, des systèmes experts, du traitement du langage naturel et de la programmation robotique.

Un projet d'IA est comme un projet d'objets connectés : il va devoir réunir des talents et compétences très divers, certaines internes aux entreprises, d'autres externes. La compétence métier prime. Suit la compétence IT plus traditionnelle, pour la collecte et l'exploitation des données.



Le paramétrage des moteurs d'IA passe par des spécialistes d'un nouveau genre qui ont de bonnes bases en IA sachant que la France en forme à peine un millier par an actuellement. Ils sont complétés par des "data scientists" qui jouent parfois tous les rôles.

Après avoir rencontré les pires des difficultés à recruter de bons développeurs, les entreprises de services, les éditeurs de logiciels, les entreprises utilisatrices tout comme les startups vont rencontrer de grandes difficultés à identifier les bons talents à même de paramétrer un moteur de deep learning[2094] !

[2093] Voir What machine learning means for software development de Ben Lorica et Mike Loukides, juillet 2018.

[2094] **Kaggle** (2010, USA, $16M), acquis par Google en mars 2017, gère une communauté de data scientists qui lance des défis aux participants. Elle génère des classements qui permettent ensuite d'identifier les meilleurs talents du marché.

Combien de temps faut-il pour apprendre à paramétrer un réseau de neurones de deep learning ou un système de machine learning ?

Il n'existe pas de réponse précise à cette question. Les cursus de formation actuels correspondent à des profils scientifiques BAC+5 avec au moins une à deux années de spécialisation.

On doit pouvoir mettre à niveau de bons profils de développeurs en moins de temps. Ceux-ci ont l'habitude de s'approprier par eux-mêmes de nouvelles techniques et outils. Les plus doués doivent pouvoir s'y mettre de manière expérimentale en quelques mois.

Il faut aussi pouvoir intéresser les salariés des entreprises qui ont été formés initialement à l'IA mais ne l'ont pas mise en œuvre en pratique car ce n'était pas à la mode au moment de leur arrivée sur le marché du travail.

Les entreprises de services numériques qui maîtrisent ce genre de projet ne sont pas encore nombreuses. Elles sont en train de s'y mettre. Aux USA, on voit se développer un peu plus qu'en France le marché des experts freelances en IA. C'est ce que propose la startup **Experfy** (2014, USA, $1,5M) avec une base de plus de 30 000 freelances.

De leur côté, les startups ne sont pas forcément adaptées à la conduite de projets, sauf pour gagner les premiers clients en entreprise. Pour les repérer, on peut commencer par visiter leurs sites web et inventorier leurs représentants, chefs de projets et ingénieurs, qui s'expriment dans les conférences sur l'IA. Les grands acteurs du service vont probablement faire l'acquisition de petits acteurs spécialisés dans l'IA pour étoffer leurs équipes.

Les projets peuvent être vite coûteux s'il faut mettre en branle une armée de consultants, data scientists, développeurs et aussi designers. Même si le cœur du réacteur d'un projet d'IA est spécifique à l'IA, autour, il faudra aussi faire tourner des briques plus classiques, tant côté back-end (préparation des données, bases de données, stockage, infrastructure, cloud) que du front-end (créer de belles interfaces pour les utilisateurs).

Où se former à l'IA ? On commence à avoir l'embarras du choix ! Les premières formations sont générales pour faire un tour d'horizon de l'état de l'art de l'IA. C'est ce que je fais chez **CapGemini Institut** depuis fin 2017 (synopsis). Ce même organisme de formation propose un cursus technique sur le machine learning (synopsis). D'autres formations à l'IA sont proposées par la **Cegos** (pour créer un chatbot, synopsis), **Orsys** (big data et intelligence artificielle avec un parcours complet pour les data scientists, synopsis), le **CNAM** (synopsis), **CNRS Formation** (état de l'art de l'IA, synopsis et un catalogue très riche sur le machine learning, le deep learning, le traitement de l'image), **Comundi** (impact de l'IA sur l'entreprise, synopsis)**, CentraleSupelec Exed** (enjeux et technologies de l'IA, synopsis), **CapDigital** (avec un parcours pour les PME auquel de je participe pour une courte session sur l'IA, synopsis), l'**Insead** (avec des cursus pour les décideurs, comme AI for business, synopsis), **IA²** (l'Institut d'Automne en Intelligence Artificielle qui propose des formations réalisées par des chercheurs). Il existe des formations longues comme celles de l'**EISTI** (synopsis) ou de **Simplon.co** (synopsis).

Dans les métiers techniques, on peut aussi choisir de se former par soi-même avec les nombreuses ressources du web. Les cours des grandes universités américains sont souvent en ligne (**Stanford**, **Cornell**, etc). On peut se former en ligne sur les réseaux de neurones et le deep learning sur **Fast.ai**, chercher des cours sur **Coursera** ou sur **DataCamp** (100 cours pour le machine learning et la data science).

Enfin, citons quelques communautés et événements de l'IA avec les **JFPC** (Journées Francophones de la Programmation par Contrainte), Journées d'Intelligence Artificielle Fondamentale, **JFPDA** (Journées Francophones sur la Planification, la Décision et l'Apprentissage pour la conduite de systèmes), **JFRB** (Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes), **JFSMA** (Journées Francophones sur les Systèmes Multi-Agents), les meetups **Machine Learning Applications Group** (Paris, Nantes, Lyon, Pau, Rennes, Aix-Marseille, Bordeaux…).

Et puis bien entendu, pour les férus de science, les deux conférences mondiales annuelles de référence, **IJCAI** et **NIPS**. Il faut ajouter **ICCV**, la conférence sur la vision artificielle. L'édition 2019 qui se tenait à Séoul attirait 7500 participants.

Startups

Les grandes entreprises font appel aux startups de l'IA pour faire évoluer leurs systèmes d'information. Elles ont maintenant l'embarras du choix avec plus de 300 startups françaises et toutes les filiales établies d'acteurs étrangers, surtout américains.

Nombre de startups de l'IA sont des sociétés de service outillé, dotées de divers outils de création de solutions et travaillant en mode projet avec les entreprises. J'étais ainsi étonné du nombre de startups dans la vision artificielle qui exposaient dans le salon AI Paris en juin au Palais des Congrès. Elles utilisent généralement des techniques courantes qui associent différentes formes de réseaux de neurones convolutifs. Leur mise en œuvre dépend ensuite du métier du client sauf lorsque la solution peut-être générique comme pour la vidéosurveillance de personnes.

Petit à petit, le marché des startups de l'IA se confond avec celui des éditeurs de logiciels. En effet, dès lors qu'un logiciel gère des volumes de données qui grandissent avec le temps, il devient possible d'exploiter ces données avec du machine learning pour faire des prévisions, de la segmentation ou de la labellisation de situations ou objets divers.

Organisation

La tentation est grande de créer de nouvelles fonctions de direction autour de l'IA. Après le Chief Digital Officer, le Chief AI Officer ? Pas forcément[2095] !

Les équipes existantes peuvent et doivent s'emparer de l'IA :

- Les **DSI** pour intégrer leur dimension technique, l'urbanisation du système d'information, le lien entre les applications de l'IA et le legacy, la gestion des données (data-lake, etc) et les infrastructures.
- Les équipes de **maîtrise d'ouvrage**, où qu'elles soient, pour faire le croisement des potentialités technologiques de l'IA avec les besoins et priorités de l'entreprise.
- Les **CDO** car l'IA est très utile dans les outils associés à leur mission, en particulier dans tout ce qui touche au marketing et à la relation clients.
- Les **business units**, qui doivent être sensibilisées aux bonnes pratiques autour de l'IA dans leur secteur d'activité.
- Les **équipes d'innovation ouverte** qui doivent faire de la veille sur les applications et techniques de l'IA comme d'autres domaines et identifier notamment des startups intéressantes pour les métiers de l'entreprise.
- Les **équipes juridiques** qui doivent être mises dans la boucle lorsque des données personnelles sont en jeu dans les applications de l'IA.

Nommer un Directeur de l'IA serait l'équivalent de nommer un « Directeur du Logiciel » tant l'IA va devenir omniprésente[2096]. Ce qui n'empêche pas nombre de grandes entreprises de créer de petites équipes de projets pilotes dans l'IA (chez Thalès, Valéo, Generali, EDF avec son programme Aria, etc) avant que l'IA ne devienne monnaie courante dans les équipes métiers, informatiques et « digitales » (marcom).

[2095] Point de vue partagé dans Qui doit gérer la stratégie en entreprise, de Robin Ferrière, Orange Business Services, septembre 2017.

[2096] Voir Qui doit gérer la stratégie d'IA en entreprise, de l'Atelier BNP Paribas, septembre 2017.

Ethique

Les nombreux appels internationaux à la prise en compte de l'éthique dans le déploiement de l'IA ne doivent évidemment pas tomber dans l'oreille de sourds dans les entreprises. Cela va notamment concerner l'épineuse question des biais de l'IA engendrés en particulier par les données d'entraînement. Tout dépendra évidemment du domaine d'application de l'IA.

Dans ce domaine, l'agence de notation française **Vigeo Eiris** lançait en 2019 une notation de prise en compte de l'impact de l'IA par les entreprises. Au sens : est-ce que le principe de précaution et les normes internationales de l'IA sont bien prises en compte ?

Juridique

Le déploiement de solutions d'IA par les entreprises doit être également examiné sous l'angle juridique et éthique. Se posent de nombreuses questions liées à la protection de la vie privée que nous avons déjà en partie évoquées ainsi que d'autres qui ont trait à la propriété intellectuelle et au droit des affaires.

Le droit des affaires de l'IA est encore un peu flou. Nombre de juristes travaillent sur l'application du droit existant aux applications de l'IA et à son adaptation éventuelle. Ces débats ont lieu un peu partout dans le monde. Cependant, le droit actuellement en vigueur permet de couvrir un bon nombre de cas d'usages de l'IA. Malgré tout, il risque d'évoluer au niveau communautaire.

Voici quelques unes des questions à se poser, sachant que vous ferez appel à vos conseils juridiques et de propriété intellectuels habituels pour en avoir le cœur net !

Responsabilité civile

Les questions en suspens portent sur la responsabilité civile de l'IA. Comment est-elle partagée entre le créateur et l'exploitant ? Cette question va se poser en particulier pour les véhicules et tous les systèmes autonomes.

Les responsabilités sont multipartites lorsque l'IA est créée par une entreprise et les données qui l'alimentent par une autre, le tout étant exploité par une troisième entité. En cas de problème, on se repose sur les responsabilités humaines qui peuvent être identifiées a posteriori en cas de différent juridique[2097].

Brevets

Il est difficile de breveter un algorithme en général et celui d'une IA en particulier. Il se trouve par ailleurs que la plupart des algorithmes utilisés dans les applications d'IA sont monnaie courante et pas forcément originaux. Pourtant, des brevets intégrant de l'IA sont régulièrement déposés. Ils utilisent les mêmes contournements que les brevets logiciels en Europe qui ne sont pas possibles sur le papier et sont pourtant nombreux. Les originalités doivent être situées ailleurs que dans les algorithmes à proprement parler.

Quid des inventions créées par les IA ? C'est pour l'instant une question théorique. Le plus souvent, les créations artistiques ou autres créations issues de l'IA sont le résultat de la manipulation d'outils de l'IA par des humains. Ce sont eux qui peuvent obtenir la paternité de ces inventions ou créations au même titre que le graphiste est bien le créateur de ses dessins et pas les outils logiciels qu'il a exploités voir combinés pour les réaliser, aussi sophistiqués soient-ils.

[2097] Voir Questions juridiques au sujet de l'intelligence artificielle par Marie Soulez du cabinet Lexing Alain Bensoussan Avocats, 2017 (5 pages) et la partie consacrée aux questions juridiques du plan France IA de mars 2017 : Intelligence Artificielle Enjeux Juridiques (19 pages). Insiste notamment sur la notion de responsabilité partagée dans la création de robots entre le matériel et les logiciels, qui ne proviennent pas des mêmes sociétés.

Il en va de même pour le créateur industriel qui exploite des outils de conception assistée par ordinateur ou le créateur de musique qui exploite un séquenceur MIDI.

A noter cet avis de l'OMPI ou WIPO, l'organisation mondiale de la propriété intellectuelle qui évoque la création d'outils à base d'IA pour accélérer la validation des demandes de brevets et de marques, notamment pour améliorer les recherches d'antériorité[2098].

Secret industriel

Le secret industriel est tout à fait intéressant pour protéger des créations à base d'IA, surtout pour des systèmes d'IA où l'entraînement voir l'exécution des traitements ont lieu de manière protégée dans le cloud. Il est alors difficile d'en faire le *retroengineering* !

Dans ce cas, l'entreprise doit cependant se protéger contre les vols de procédés ou données réalisées par ses propres salariés ou sous-traitants.

Droit d'auteur

Le droit d'auteur s'applique aux logiciels et aux créations de l'IA comme toutes les œuvres déjà créées par l'Homme avec l'assistance de machines. Seule une personne physique est protégée par le droit d'auteur. La machine 100% autonome et créatrice n'existe pas encore[2099]. Par contre, on va pouvoir utiliser des IA pour identifier de manière plus rapide d'éventuels plagiats.

Les droits d'auteurs devraient notamment pouvoir s'appliquer à des modèles de deep learning IA entraînés. A tel point que des chercheurs d'IBM ont même développé une méthode de watermarking de modèles de deep learning[2100] ! C'est en gros un moyen d'intégrer une signature dans un modèle de deep learning pour vérifier qu'il est bien d'origine lorsqu'il est exploité. C'est une méthode voisine de celle qui protège des contenus protégés par le copyright.

Données

A qui appartiennent les données et les modèles entraînés par le machine learning ou le deep learning ? La question est pour l'instant la même qu'avec les données exploitées dans des applications traditionnelles qui n'exploitent pas d'IA.

L'autre question juridique clé à traiter est le respect du RGPD. Il impose de bien se documenter sur la manière d'anonymiser les données dans l'apprentissage d'applications d'IA. Il peut être intéressant de creuser la manière d'exploiter la *differencial privacy* qui assure que des modèles entraînés ne permettront pas de reconnaître des utilisateurs individuels après requêtage[2101]. Il est cependant rare qu'une application à base d'IA ne puisse pas être lancée à cause du RGPD.

Contrats

Les entreprises vont faire appel à de nombreux sous-traitants pour créer des solutions d'IA. Quelle est la propriété intellectuelle attachée à ces solutions lorsqu'il y a eu une forte personnalisation de solutions ? Il va falloir faire la part des choses entre contrat de service et contrat de licence d'utilisation de logiciels ou la combinaison des deux.

[2098] Voir Artificial intelligence and intellectual property: an interview with Francis Gurry, septembre 2018.

[2099] Voir La protection par le droit d'auteur des créations générées par intelligence artificielle, un mémoire de Maîtrise en droit de Claudia Gestin-Vilion de l'Université Laval de Québec et de l'Université Paris-Saclay à Sceaux, 2017 (112 pages).

[2100] Voir Protecting Intellectual Property of Deep Neural Networks with Watermarking, IBM Research, 2018 (13 pages) ainsi que DeepMarks: A Digital Fingerprinting Framework for Deep Neural Networks, 2018 (12 pages).

[2101] Voir Le RGPD bride t-il l'IA ? par Grégory Herbert de Dataiku, juin 2019.

Epilogue

Nous voici au terme de ce voyage à 360° dans l'IA qui se voulait aussi pratique et pragmatique que possible et destiné avant tout aux entreprises qui se demandent par quel bout prendre le sujet de l'intelligence artificielle.

Vous aurez saisi que les prouesses récentes de l'IA sont liées à des progrès parallèles : dans les méthodes et algorithmes qui s'améliorent continûment, dans les données qui les alimentent et dans le matériel. La puissance brute des machines ne fait pas tout, même si elle peut avoir tendance à rendre les développeurs moins astucieux dans leur manière d'aborder les problèmes. Cette impression vient de la difficulté à appréhender la nature même des progrès réalisés dans les algorithmes de l'IA car leur vulgarisation est très difficile. Si vous avez bien saisi comment fonctionnait un réseau de neurones convolutionnel et ses applications, vous avez déjà franchi une belle étape ! Reste à faire de même pour les réseaux de neurones de traitement du langage, ce qui est moins évident.

Vous avez aussi pu constater que l'IA est encore un secteur assez artisanal, aussi bien chez les chercheurs que chez les startups. Elles assemblent des briques techniques de manière encore très expérimentale. Les startups du logiciel de l'IA font encore surtout du service outillé. L'évolution de l'industrie du logiciel de l'IA vers une véritable approche produit est un parcours semée d'embûches. Seules les entreprises b2c y arrivent véritablement pour l'instant, grands acteurs en premier.

Malgré ces nombreux écueils et la bulle médiatique qui l'accompagne, la vague de l'IA est sérieuse et semble aussi importante que les vagues technologiques précédentes qu'ont été le cloud, le big data ou la mobilité. L'IA est en quelque sorte une vague « 2.0 » du logiciel. C'était la *Belle au Bois Dormant* pendant près de 60 ans, parallèlement avec l'Histoire de l'informatique traditionnelle. Elle s'est éveillée depuis une douzaine d'années pour ne plus se réendormir. Aucune sorcière ne l'en empêchera.

Nous avons pu découvrir de nombreuses startups françaises et étrangères, aussi bien au niveau des techniques horizontales que des applications métiers. Nous ne manquons pas de chercheurs et d'entrepreneurs de l'IA en France. Les questions qui se posent sont les mêmes que d'habitude : comment faire en sorte que ces startups grandissent vite, soient bien financées et se développent à l'international. Les startups et grands acteurs du numérique US et chinois ne nous ont pas attendus !

Même avec une IA approximative, la marche vers l'automatisation partielle de nombreux métiers est déjà en route et va dans le sens d'une histoire qui a démarré avec l'invention de la roue. L'IA est une nouvelle boîte à outils à intégrer dans son atelier logiciel. Il faut s'y préparer dès maintenant, ne pas y résister futilement, s'y adapter en se modernisant, en faisant évoluer notre système d'enseignement et en produisant des outils compétitifs.

Les civilisations qui ont évité les progrès techniques et les outils de communication dans l'Histoire ont systématiquement périclité ou, au mieux, décliné. Le futur n'est pas écrit à l'avance, il s'écrit au fur et à mesure par les innovateurs. C'est la société qui adopte ou pas les innovations en fonction de motivations complexes. L'équilibre qui se construit est le résultat de choix collectifs complexes.

Aux entreprises donc de se moderniser et de créer des solutions qui, certes, répondent à des besoins métiers et exploitent l'IA, mais aussi de le faire avec responsabilité, avec les bons garde-fous pour éviter des dérives que l'on découvre, que ce soit au niveau du respect de la vie privée ou du simple besoin de relations humaines que nous pouvons toujours ressentir. Réduire à outrance les relations humaines sous couvert d'efficacité économique n'est pas ce à quoi l'Homme aspire naturellement. Donc, une IA outil, oui, mais pas une IA qui nous fait nous passer d'Humanité.

Médias spécialisés

L'intelligence artificielle est devenue un thème couramment couvert par l'ensemble de la presse technologique, scientifique et économique généraliste. Il n'existe pas beaucoup de médias spécialisés sur l'IA en dehors de revues pointues destinées aux chercheurs.

En voici quelques exemples.

2051.fr : est un site d'information sur l'IA plutôt tourné vers les besoins des entreprises.

A.I. Magazine : un site d'information en anglais sur l'IA. ai-magazine.com.

ActuIA : un site de news en français sur l'IA, www.actuia.com/acteur/list

AI Magazine : un magazine US sur la recherche en IA édité par l'Association for the Advancement of Artificial Intelligence. www.aaai.org/Magazine/magazine.php.

AI Playbook : un site du fonds d'investissement Andreesseen Horowitz qui défriche le champ d'application de l'IA dans les entreprises. aiplaybook.a16z.com/

Arxiv : le site de publication de papiers de chercheurs coordonné par Cornell University, avec sa version dédiée à l'IA comprenant plus de 57 000 documents, www.arxiv-sanity.com/, dont la curation est assurée par Andrej Karpathy.

Chatbot Magazine : un site US sur les chatbot. chatbotsmagazine.com/.

In Principio, un site de vulgarisation sur l'IA couplé à un blog d'actualité, www.inprincipio.xyz.

Journal of Intelligence Artificial Research : qui comment son nom l'indique couvre l'actualité de la recherche en IA. www.jair.org/.

Journal of Machine Learning Research, www.jmlr.org.

Mais où va le web : qui commente l'actualité du numérique, dont celle de l'IA, avec une vision critique et caustique. maisouvaleweb.fr/.

Nanalyze est un bon site web faisant le tour de l'actualité autour des startups de l'IA. www.nanalyze.com.

Off the convex : un bon blog sur le deep learning, http://www.offconvex.org

Singularity Hub : magazine de l'actualité scientifique teinté par les technologies dites exponentielle, singularityhub.com/

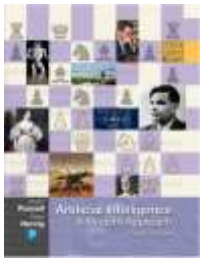
Topbots : un site d'actualité sur l'état de l'art de l'IA, notamment dans la recherche et qui publie des synthèses de publications de chercheurs très intéressantes (exemple). www.topbots.com.

Voicebot : un autre magazine US sur les chatbots. www.voicebot.ai/.

Wonk : un site de l'OCDE qui fait intervenir des experts internationaux de l'IA.

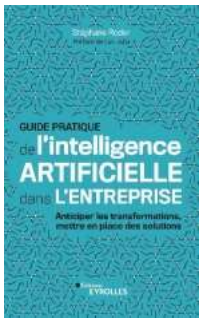
Bibliographie

Voici quelques ouvrages, que je n'ai pas forcément tous lus, sur l'IA. L'édition est très abondante sur le sujet depuis des années. On y trouve de tout, et surtout des ouvrages de vulgarisation destinés soit aux entreprises soit au grand public. Certains présentent une vision inquiétante de l'IA, d'autres optimistes. On peut identifier divers biais chez nombre d'auteurs en consultant leur biographie.



Une des manières d'évaluer l'évolution du champ scientifique de l'IA est de se se plonger dans l'ouvrage de référence **Artificial Intelligence A Modern Approach**, de Stuart Russell et Peter Norvig. Sa quatrème et dernière édition d'avril 2020 fait la bagatelle de 2145 pages pour 1152 pages dans l'édition précédente de 2010. Elle a été longtemps l'un des B-A-BA pour les étudiants en informatique et en IA de premier cycle[2102]. Elle donne la part belle aux agents, aux techniques de recherches apparentées à la force brute utilisée notamment dans les jeux de société, à la programmation par contraintes, au raisonnement avec les agents logiques, la logique du premier ordre, la planification, la représentation des connaissances et les méthodes probabilistes et bayésiennes. La nouvelle édition de 2020 intègre le deep learning, le traitement du langage et de la robotique.

Quand la machine apprend par Yann Le Cun, octobre 2019 (394 pages) est le dernier ouvrage du plus fameux scientifique français de l'IA, responsable de la recherche en IA chez Facebook au niveau mondial. Il décrit les avancées de la branche de l'IA qui est sa spécialité, le deep learning. La partie pédagogique sur l'histoire et le fonctionnement des réseaux convolutifs est très intéressante. Il semble que pour lui, ceux-ci soient bons à tout faire. Ses travaux de recherche, déjà évoqués à différents endroits dans cet ebook, visent à perfectionner le deep learning pour lui permettre d'apprendre à la volée et raisonner de manière symbolique. Il défend ainsi l'émergence future d'une IA au niveau de l'intelligence humaine tout en tempérant les performances des IA d'aujourd'hui qui sont survendues comme celles du robot Sophia d'Hanson Robotics[2103].

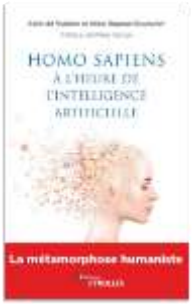


Guide pratique de l'intelligence artificielle dans l'entreprise de Stéphane Roder, 2019 (208 pages) souhaite aider à définir la stratégie d'intelligence artificielle des entreprises. L'auteur adopte une approche utilitariste et économique de l'IA pour en décrire les bénéfices business et aider à identifier d'éventuels gains de productivité et de croissance liés à l'exploitation de l'IA. C'est accompagné d'une méthodologie de mise en œuvre des solutions d'IA dans les entreprises. L'auteur est le fondateur de la société de conseil AI Builders en 2018 qui accompagne les entreprises dans l'adoption de l'IA. Il est aussi enseignant à l'ESSEC. Il a heureusement un background d'ingénieur télécom et une véritable connaissance des technologies de l'IA qui évite de raconter n'importe quoi comme cela peut arriver dans les métiers du conseil. L'auteur est aussi référencé comme expert auprès de Bpifrance.

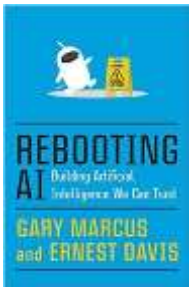
La révolution des services 4.0, de John et Matt Rauscher**,** 2019 (194 pages) est un livre destiné aux dirigeants de PME qui leur permettra d'identifier les usages de l'IA dans leur quotidien business. Il fait bien la part des choses entre le machine learning et le raisonnement automatique et fournit une vingtaine d'exemples d'usages de l'IA dans l'entreprise notamment pour le *knowledge worker* et la gestion de la relation client. John Rauscher est le fondateur de la société Yseop, spécialisée dans le traitement du langage. Les auteurs expliquent pourquoi l'IA pourrait créer de l'emploi.

[2102] Le Russell et Norvig 2010 est téléchargeable gratuitement ici. Ses auteurs prévoyaient en 2017 de sortir une nouvelle édition en 2019 qui devait couvrir notamment les avancées du deep learning. Le projet a visiblement pris du retard !

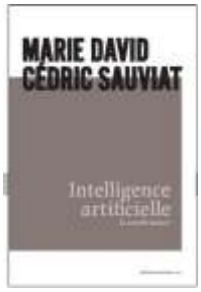
[2103] Voir Les machines vont arriver à une intelligence de niveau humain, octobre 2019.



Homo sapiens à l'heure de l'intelligence artificielle par Alain de Vulpian (sociologue et anthropologue) et Irène Dupoux-Couturier (historienne), 2019 (240 pages) développe la thèse d'un véritable tournant du destin de l'espèce humaine provoqué par l'IA. Ils adoptent une approche sociétale de l'évolution de l'espèce humaine face aux bouleversements technologiques qu'ils qualifient de métamorphose radicale. Homo sapiens mettrait « *en dialogue ses intelligences rationnelle, émotionnelle relationnelle, sensorielle et spirituelle* ». Face à de telles thèses, je ne peux m'empêcher de relativiser l'importance de l'IA en tant que telle et de l'intégrer dans le cadre plus général des transformations de la vie quotidienne générée par l'ensemble des usages du numérique. L'IA n'en est qu'un des aspects, sauf à l'élever à une position mythique assez éloignée de la réalité des usages d'aujourd'hui.



Rebooting AI – Building Artificial Intelligence We Can Trust de Gary Marcus et Ernest Davis, 2019 (290 pages) est un ouvrage qui casse de nombreux mythes de l'IA et la remet à sa juste place, celle d'une boîte à outils, comme décrite dans cet ebook. Les auteurs sont des psychologues. Ils rappellent le caractère étroit et spécialisé des IA d'aujourd'hui, en commençant par celles qui gagnent aux jeux comme AlphaGo. Ils montrent que les IA connexionnistes d'aujourd'hui ne comprennent rien à ce qu'elles font qu'il s'agisse de la reconnaissance d'images ou le traitement du langage. Bref, nous sommes aujourd'hui face à des moteurs probabilistes qui peuvent accompagner le travail des humains mais pas le remplacer. Ils militent pour une IA plus douée dans le raisonnement, bref la fameuse « IA symbolique » un peu délaissée depuis l'avènement du deep learning ainsi que pour des IA auditables et explicables.



Intelligence artificielle, la nouvelle barbarie par Marie David et Cédric Sauviat, 2019 (320 pages) est issu de deux auteurs polytechniciens qui s'inquiètent des dérives potentielles de l'IA notamment en termes d'impact sur la vie privée et sur l'influence de notre vie quotidienne. Ils dénoncent le risque de la passivité de l'Homme face à l'IA, et, j'imagine, aux abus des outils numériques. En découleraient des menaces graves sur la démocratie. Ils considèrent que l'IA n'est pas une technologie numérique comme les autres et qu'il faut donc l'adopter en utilisant une sorte de principe de précaution. Le message porté par le livre s'explique indirectement par le fait que Cédric Sauviat est en fait le président de l'AFCIA, l'Association Française contre l'Intelligence Artificielle[2104] !

Guerres économiques pour l'intelligence artificielle, par Benoît Berard, Clovis Fayolle et Benoît Pahud, mai 2019 (386 pages) traite des questions géopolitiques et de souveraineté de l'IA. Les auteurs sont des spécialistes de la stratégie des entreprises, pour l'essentiel dans l'industrie. Malheureusement, la trace qu'ils ont laissé sur Internet est bien pauvre, avec notamment, aucune vidéo de référencée sur YouTube.

La face cachée de l'intelligence artificielle par Boussad Addad, septembre 2020 (234 pages) qui couvre la géopolitique de l'IA et la guerre économique associée, surtout entre la Chine et les USA. Il traite notamment des enjeux concernant les semi-conducteurs mais aussi l'affaire Cambridge Analytica. Cette présentation de défis sociétaux et politiques posés par l'IA se termine par une fiction de prospective.

[2104] Voir Pourquoi résister à l'Intelligence artificielle ? par Cédric Sauviat, mai 2018.

La révolution des assistants vocaux par Oxana Gouliaeva, Eric Dosquet et Yvon Moisan, janvier 2020, avec une préface de Luc Julia (236 pages). L'ouvrage est un guide pratique destiné en premier lieu aux marketeurs. Il décrit la problématique et les domaines d'applications des assistants vocaux. Il contient nombre de témoignages et études de cas. L'ouvrage n'est pas technique mais vraiment orienté usages et business. On y trouve notamment des parties concernant les usages des assistants vocaux le long du parcours client, la structure du marché et la gestion de projets.

Dictionnaire anglais/français de l'IA

Anglais	Français	Commentaire
ASR – Automatic Speech Recognition	Reconnaissance de la parole.	Intégré dans les agents conversationnels vocaux.
Back propagation	Rétro-propagation	Technique d'apprentissage de réseau de neurones
Capsule networks	Réseaux à capsules	Créés par Geoff Hinton fin 2017.
Convolutional networks	Réseaux de convolution, réseaux convolutifs	Type de réseau de neurones pour le deep learning.
Deep learning	Apprentissage profond	Machine learning avec réseaux de neurones à grand nombre de couches.
Differential privacy	Confidentialité différentielle	Méthode d'anonymisation des données et logiciels, dont ceux de l'IA.
Embedding	Enchassement	Représente un objet (mot, image) par un vecteur, sorte de hash code.
Feature map	Carte de caractéristiques.	Utilisée dans les convnets
Feed forward neuron network	Réseau de neurones avec réglage en avant	Intraduisible correctement !
Filters	Filtres	Utilisée dans les convnets
Fonction de coût	Cost function	Utilisée pour l'entraînement de modèles de machine et de deep learning.
Generative Adversarial Network	Réseau génératif antagoniste	Utilisé pour créer des contenus à partir d'autres contenus.
GDPR : General Data Protection Regulation	RGPD : Règlement Général de Protection des Données	Nouvelle régulation européenne de protection des données personnelles applicable depuis mai 2018.
Homomorphic cryptography	Chiffrement homomorphe	Utilisable pour faire gérer des données chiffrées dans le machine learning.
k-means	k-moyennes	Méthode de regroupement des données ou clustering dans le machine learning.
Machine learning	Apprentissage automatique	
Neuromorphic chipsets	Composants neuromorphiques	Processeurs spécialisés pour les réseaux de neurones.

Neural networks	Réseaux de neurones	Réseaux de neurones artificiels utilisés dans le machine learning et le deep learning.
Optical Characters Recognition	Reconnaissance de caractères	
Pooling	Aggrégation	Dans les réseaux de neurones convolutifs.
Prediction	Prévision (générées par le machine learning) Prédiction (pour madame Soleil et boules de cristal)	Faux ami !
Principal Components Analysis	Analyse en composantes principales	Une technique utilisée dans le machine learning.
Quantum computing	Informatique quantique	A des applications dans le machine learning.
Recurrent neuronal networks	Réseaux de neurones récurrents	Type de réseau de neurones pour le deep learning.
Reservoir Computing	*(peu traduit)*	Technique de réseau de neurone à base d'unités récurrentes.
Stockastic Gradient Descent	Descente stockastique de gradient	Utilisée dans la back-propagation
Shallow networks	Réseaux à faible profondeur	Utilisé dans le machine learning.
Sparse	Parcimonieux	Type de réseaux de neurones.
Spiking neurons	Neurones à impulsions, neurones impulsionnels	Utilisés notamment dans le traitement du langage et dans certains processeurs neuromorphiques.
Stacked autoencoders	Autoencodeurs empilés	Réseaux de neurones générant des contenus à partir de descripteurs.
SVM : support vector machines	Machine à vecteurs de support, séparateurs à vastes marges	Technique de segmentation du machine learning.
Tree search	Arbres de décision	Technique de machine learning pour la classification ou la régression.
Transfer learning	Apprentissage par transfert	Apprentissage d'un réseau de neurones incrémental ou appliquant la reconnaissance d'un domaine d'application à un domaine voisin.

Uncanny valley	Vallée de l'étrange	Phénomène se manifestant lorsque l'on est mal à l'aise face à un robot humanoïde trop réaliste.
Variational Auto Encoder	*(peu traduit)*	Utilisés dans les réseaux de neurones génératifs.

Glossaire

AGI : Artificial General Intelligence, IA de niveau équivalent à celle de l'Homme. Tout du moins dans la capacité de raisonnement. C'est un Graal pour certains chercheurs et singularistes qui croient qu'il suffira d'attendre les effets de la loi de Moore pour y arriver. Les protagonistes de l'AGI oublient presque systématiquement d'indiquer à quoi celle-ci servirait et le genre de problème que l'on pourrait lui soumettre.

Alexa : service en ligne d'agent conversationnel d'Amazon, fonctionnant par reconnaissance vocale et intégré dans ses objets connectés de la gamme Echo et des objets connectés tierce-partie créés par les partenaires d'Amazon.

Algorithmes génétiques : algorithmes s'améliorant d'eux-mêmes par un processus d'évolution voisin de celui du vivant, avec des techniques de croisements et de sélection darwinienne.

ANI : Artificial Narrow Intelligence, IA utilisée dans un champ précis de résolution de problèmes. C'est l'état de l'art actuel.

Apprentissage : se dit du réglage des paramètres d'un modèle de machine learning ou de deep learning. Il est soit supervisé avec des données d'entraînement labellisées soit non supervisé, pour ce qui est de l'identification de classes d'objets différentiées.

Apprentisssage par tranfert / transfer learning : apprentissage d'un modèle de deep learning exploitant celui d'un réseau existant. C'est une sorte de processus d'apprentissage incrémental qui éviter le fastidieux réentraînement complet d'un réseau de neurones si on ajoute une nouvelle classe d'objets à reconnaître.

ASI : Artificiel Super Intelligence, IA de niveau supérieure à celle de l'Homme. Elle serait la conséquence immédiate de l'AGI qui se reproduirait et de démultiplierait d'elle-même.

ASIC : circuits intégrés intégrant des portes logiques gravées en dur. Les chipsets de mobiles et les microprocesseurs sont des ASIC. Ils présentent l'avantage de consommer moins d'énergie et d'être plus rapides que les FPGA mais ne sont intéressants économiquement que s'ils sont produits en grand volume. Technique utilisée par Google pour ses processeurs neuromorphiques TPU.

Back propagation : rétro-propagation, technique d'entraînement de réseau de neurones consistant à comparer le résultat du réseau sur un objet type avec la bonne classe de l'objet et de rétropropager l'erreur en remontant dans le réseau de neurones. Cela utilise des gradients, des fonctions de couts et plein de concepts divers et variés. Cette rétropropagation est réalisée pour tous les objets de la base d'entraînement. C'est un traitement très couteux en ressources machines. Il est possible de l'automatiser pour le paralléliser sur des architectures multi-cœurs ou multi-processeurs. Il est encore plus efficace dans les processeurs neuromorphiques.

Bayésien : technique d'IA s'appuyant sur des modèles probabilistes et statistiques.

BRMS : Business Rules Management Systems, les logiciels de gestion de règles permettant de créer des systèmes experts.

Caffe : bibliothèque de développement de solution de machine learning et de deep learning pour la mise en production. L'outil open source provient à l'origine de l'Université de Berkeley et est promu par Facebook, son créateur Yangqing Jia travaillant chez eux. La version courante est la 2.0. En avril 2018, Caffe2 a été intégré dans PyTorch.

Classification : se dit des modèles de machine learning et de deep learning qui servent à associer un objet à une classe d'objet. Comme une photo à un nom d'objet contenu dans la photo.

CNN : Convolutional Neuronal Networks, ou réseaux de neurones convolutifs.

Connexionnisme : méthode et techniques de l'IA mettant en œuvre une modélisation à bas niveau à base de réseaux de neurones artificiels.

ConvNet : Convolutional Neuronal Networks, ou réseaux de neurones convolutifs.

Convolution : couche de convolution, étape d'un réseau de neurones convolutionnel qui détecte des formes dans une image. Un réseau convolutif contient plusieurs couches de convolution qui détectent des formes aant un niveau d'abstraction de plus en plus élevé au gré de l'avancement du traitement dans le réseau.

Cortana : agent conversationnel de Microsoft.

Cybernétique : étude du fonctionnement des systèmes complexes s'appuyant sur l'échange d'information, élaborée en 1947 par Norbert Wiener. Les travaux associés se sont étalés entre 1942 et 1953. Ils s'appuient sur la notion de boucle de feedback qui permet d'expliquer aussi bien le fonctionnement du vivant que des machines. On en retrouve les concepts dans les systèmes multi-agents ainsi que dans l'apprentissage de réseaux de neurones par descente de gradient.

DARPA : agence américaine de financement de la R&D pour le Pentagone. L'un des plus grands financeurs de projets de R&D dans l'IA au monde. Elle finance sur challenges et appels à projet des projets qui sont réalisés par des laboratoires privés et publics, des startups, PME et grandes entreprises.

Datalake : petit nom des entrepôts de données des entreprises qui servent ensuite au développement d'applications de machine learning et de deep learning. Ils stockent en général des données aussi brutes que possible, sans préjuger trop à l'avance de leur usage futur.

Sachant qu'elles doivent cependant être aussi propres que possible.

DBN : machines restrictives de Boltzmann, des réseaux de neurones datant de 1986 utilisant une seule couche de neurones source et cible et sans connexions entre les neurones d'une même couche. C'est le modèle le plus simple de réseau de neurones qui est ensuite exploité dans d'autres assemblages, comme les Deep Belief Networks (DBN) créés en 2006.

Decision Management Systems : concept marketing de système d'optimisation des décisions des entreprises qui englobe les moteurs de règles pour codifier les connaissances humaines et pratiques de l'entreprise, des modèles prédictifs qui utilisent le machine learning pour recommander les actions, et des outils d'analytics de reporting.

Deep Blue : nom de l'ordinateur qui a gagné aux échecs contre Gary Kasparov en 1997. Il s'agissait en fait d'un modèle avancé, dénommé initialement Deeper Blue.

Deep learning (apprentissage profond) : extension du machine learning intégrant des fonctions d'apprentissage supervisé et d'auto-apprentissage s'appuyant sur des modèles de représentation de données complexes et multi-dimensionnels. Ce sont des réseaux de neurones avec un grand nombre de couches de neurones.

Deep Mind : filiale de Google acquise au Royaume-Uni en 2014. Est à l'origine de la victoire contre le champion mondial de Go début 2016 ainsi que d'AlphaFold, la solution de deep learning qui simule le repliement de protéines..

Edge computing / **edge AI** : principe consistant à déporter près des objets et capteurs les calculs d'IA relatifs aux données qu'ils génèrent. Cela peut être le cas de centrales domestiques de gestion de caméras de surveillance. C'est un juste milieu entre la réalisation des traitements sur des serveurs d'inférences dans le cloud ou dans les objets eux-mêmes (endpoints).

Entraînement : synonyme d'apprentissage pour un modèle de machine learning et de deep learning.

Feature map : composante des réseaux de neurones convolutifs. Il s'agit d'une matrice qui contient des valeurs décrivant la pondération de l'apparition d'un filtre dans une image d'origine. Un filtre contient une forme donnée. Si celle forme est détectée, cela va donner un 1, si elle n'est pas détectée, cela donne un 0. Et toute une gradation entre zéro et un pour les valeurs intermédiaires.

Federated Learning consiste à entraîner des réseaux de neurones localement et à ne transmettre que les propriétés et paramètres des réseaux neuronaux locaux à un réseau neuronal d'un niveau supérieur. Cela peut permettre d'éviter d'exposer ses données à l'extérieur, en lieu et place ou en complément du chiffrement homomorphe.

Filtre : utilisé dans un réseau de neurones convolutionnel, sert à identifier des formes avec des niveaux d'abstraction plus ou moins élevés. Ils sont d'abord initialisés de manière aléatoire puis ajustés progressivement par entraînement du réseau de neurones avec rétropropagation des erreurs.

Fonction de coût (cost function) : fonction qui évalue évalue l'erreur de reconnaissance d'objet lors de l'entraînement par apprentissage supervisé. L'entraînement de modèles de machine learning consiste à réduire au minimum cette fonction de coût.

Force brute : technique de résolution de problème utilisant surtout la puissance des machines et des algorithmes traditionnels, quelle que soit leur efficacité. Souvent associée à des algorithmes dits exponentiels, dont le temps de calcul évolue de manière exponentielle avec la taille du problème à traiter.

FPGA : circuits intégrés intégrant des portes logiques qui sont définies par programmation. Ils sont adaptés à la fabrication de petites séries et au prototypage. Ils consomment plus d'énergie et sont plus lents que les ASIC.

GAN : Generative Adversarial Networks, technique de réseaux de neurones convolutifs inversés qui génèrent des contenus à partir d'autres contenus ou d'informations élémentaires. Voir la partie qui en décrit des usages.

GNN : Graph Neural Networks, réseaux de neurones permettant de traiter des objets « graphes », par exemple, pour les labelliser ou les gérer (GAN de GNN).

GOFAI : « Good Old-Fashioned Artificial Intelligence » qui dénomme les méthodes d'IA s'appuyant sur les méthodes symboliques comme dans les systèmes experts, en vogue jusque dans les années 1980.

Gradient : pente de la courbe de fonction d'erreur pendant les phases d'entraînement. Une descente de gradient consiste à trouver l'orientation des paramètres des réseaux de neurones permettant de faire baisser le taux d'erreur.

GRU : Gated Recurrent Units, technique introduite dans les réseaux de neurones récurrents en 2014 qui simplifie les traitements par rapport au LSTM.

HBM2 : standard de mémoire RAM qui est placé juste à côté de processeur type CPU ou GPU. Il permet un transfert très rapidement d'information entre les deux, à raison de 900 Go/s à 1,8 To/s. C'est notamment l'architecture utilisée dans les GPU V100 de Nvidia depuis 2017 puis dans les A100 lancés en 2020.

Hivers de l'IA : périodes de creux et de désaveu dans l'histoire de l'IA. Le premier hiver date de la fin des années 1970 et le second de celle des années 1980 et début 1990.

Hyperparamètres : ce sont les paramètres qui servent à définir la structure d'un réseau de neurones : le nombre et la nature des couches ainsi que le nombre de filtres et leur taille dans les couches de convolution.

IA intégrative : technique de création de solutions d'IA associant plusieurs techniques différentes (agents, moteurs de règles, réseaux neuronaux, machine learning, deep learning, bayésien, …).

Inférence : se dit de l'exécution d'un modèle de machine learning et de deep learning avec une donnée de production, comme pour la labelliser ou faire une prévision.

Infiniband : standard de communication entre serveurs dans les data centers qui repose sur de la fibre optique. Supporte des débits de quelques dizaines de Go/s. Le leader de ce marché est l'Israélien Mellanox, acquis par Nvidia en 2019.

Kill switch : métaphore du bouton d'arrêt d'urgence d'un ordinateur doué d'IA de niveau AGI ou ASI au cas où celui-ci ne serait plus sous contrôle.

Labellisation ; processus, souvent manuel, de rattachement de labels à des objets devant ensuite servir à l'entraînement d'un modèle de machine learning et de deep learning avec apprentissage supervisé.

LISP : langage de programmation d'IA utilisé dans les années 80 et 90 et notamment dans la création de systèmes experts.

Logique floue : technique d'IA créé par Lofti Zadeh dans les années 1960 et représentant l'information non pas sous forme binaire mais sous forme floue comprise entre 0 et 1. Elle est parfois utilisée dans les moteurs de règles de systèmes experts.

LSTM : long short term memory, modèle de réseaux de neurones récurrents qui intègrent bien le contexte dans lequel les éléments apparaissent de manière séquentielle. Aussi appelés réseaux de neurones à mémoire. Ils servent en particulier à interpréter le langage et à faire de la traduction automatique.

Machine learning (apprentissage automatique) : technique d'IA permettant de résoudre des problèmes de perception de l'environnement (visuel, audio, …) de manière plus efficace qu'avec les algorithmes procéduraux traditionnels. Elle s'appuie souvent sur l'usage de réseaux de neurones artificiels.

Markov, modèle de : méthode d'IA s'appuyant sur des méthodes probabilistes.

Mask CNN : réseau de neurones qui est capable de segmenter pixel par pixels les objets dans une image. C'est une déclinaison des Faster-R-CNN.

Matrices creuses : aussi appelée sparce matrice en anglais ou aussi matrices parcimonieuses en français. Ce sont des matrices utilisées dans les modèles de deep learning dont une majorité des cellules sont à zéro. Il existe des techniques d'optimisation - de type compression - du calcul les concernant, mises en place dans certains processeurs tels que le GPGPU A100 de Nvidia ou ceux de Graphcore.

Memristors : composants électroniques à mémoire persistence. On pourrait créer avec cette technique des processeurs neuromorphiques se raprochent le plus des neurones, avec une mémoire proche des unités de traitement permettant d'aller bien plus vite dans l'entraînement de réseaux de neurones.

Modèle : se dit d'un système de machine learning ou de deep learning après entraînement. Il contient des hyperparamètres qui définissent le poids des neurones, les fonctions d'activation utilisées ainsi que la nature des couches et des filtres utilisés. Ce sont en gros les variables de fonctionnement d'un logiciel de machine learning. Elles sont définies après une phase d'entraînement.

Moteurs de règles : solutions techniques permettant de mettre en œuvre des systèmes experts et exploitant des bases de prédicats (règles).

NLP : Natural Language Processing. Techniques du traitement automatique du langage.

NPU : neural processing unit. Se dit des unités de traitement de processeurs généralistes qui réalisent des traitements pour des réseaux de neurones. En général, il s'agit de multiplication et addition de vecteurs et de matrices.

Neuromorphique : se dit des processeurs neuromorphiques qui présentent la particularité d'intégrer dans leur conception des modules de calcul qui collent avec les besoins des principaux réseaux de neurones et en particulier les réseaux de neurones convolutifs. En pratique, ils comprennent des multiplicateurs de matrices et des matrices de synapses connectant des vecteurs de neurones, plus de la mémoire locale rapide.

Neurones à impulsions : ou spiking neurons, type de composant électronique simulant le fonctionnement de neurones biologiques avec une notion temporelle du flux d'information à traiter, pendant aussi bien les phases d'entraînement que d'inférences.

Non supervisé : se dit d'un apprentissage de modèle de machine learning et de deep learning avec des données non labellisées. Un tel apprentissage ne va pas deviner les labels en question. Il sert à identifier des classes d'objets et des anomalies qu'il faudra ensuite éventuellement labelliser à la main.

Paramètres d'un réseau de neurones : les variables du réseau (poids des liaisons entre neurones et contenu des filtres) qui sont définies lors de l'entraînement du réseau de neurones. A ne pas confondre avec les hyperparamètres qui définissent la structure d'ensemble du réseau de neurones. Le réseau de neurones le plus sophistiqué en 2020, G-Shard, atteignait 600 milliards de paramètres.

Pooling : technique de réduction de résolution des feature maps dans les réseaux de neurones convolutifs. Ils permettent de réduire les temps d'entraînement et de traitements dans ces réseaux.

Prévision et **prédiction** : les deux termes sont souvent utilisés de manière indifférenciée pour décrire les méthodes de détermination d'une donnée dans le futur. Les puristes considèrent qu'il faut utiliser le terme prévision pour ce qui est généré par le machine learning. Une prédiction est une annonce d'un fait futur. Une prévision est basée sur des faits antérieurs. Les prédiction sont souvent celles d'astrologues tandis que les météorologues font des prédictions. Les premières sont ésotériques et les secondes s'appuient sur des fondements scientifiques. Mais cela peut se mêler. Pour cette raison, lorsque Kurzweil anticipe l'arrivée d'une AGI en 2040, c'est plutôt une prédiction.

PyTorch : bibliothèque de développement d'application de machine learning et deep learning open source provenant de Facebook et adaptée aux travaux de recherche et de prototypage.

Quantum Machine Learning : famille d'algorithmes de machine learning exploitables sur ordinateurs quantiques.

Rapport Lighthill : rapport anglais ayant conduit au premier hiver de l'IA en 1973 après avoir constaté les progrès trop lents de l'IA faisant suite à des promesses trop ambitieuses.

Régression : modèle de machine learning simple permettant de prédire une valeur numérique dans le temps en fonction de son évolution passée. On parle de régressions linéaires et non linéaires en fonction de la forme de l'évolution dans le temps des valeurs.

Renforcement : mode d'apprentissage de machine learning et de deep learning qui passe par une boucle de rétroaction entre le modèle et son environnement. C'est un apprentissage par l'erreur, le modèle cherchant à minimiser celle-ci. S'applique notamment aux robots pour leur apprendre à attraper une pièce, à un chatbot pour l'orienter vers les meilleures réponses en fonction du contexte des questions.

Réseaux à mémoire : réseaux de neurones capables de traiter un contenu en fonction de son contexte. La variété historique est le LSTM.

Réseaux convolutifs : réseaux de neurones utilisés principalement pour la reconnaissance d'objets dans les images et utilisant plusieurs couches de neurones (pooling) de niveaux d'abstration échelonnés des composantes de l'image.

Réseaux récurrents : réseaux de neurones adaptés à la détection de formes dans des signaux récurrents comme un battement cardiaque.

Réseaux de neurones : technique d'IA visant à simuler le fonctionnement des cellules neuronales pour reproduire le fonctionnement du cerveau humain. Est surtout utilisée dans la reconnaissance de la parole et des images. Peut-être simulé en logiciel ou avec des circuits électroniques spécialisés.

Rétropropagation : des erreurs, des gradients, se dit des méthodes d'entraînement des réseaux de neurones qui passe par la propagation à partir de la fin du réseau des erreurs observées pendant les phases d'entraînement avec les données d'entraînement labellisées.

R-CNN : regions based CNN, réseau de neurones qui est capable de détourer des objets dans des images et de les labelliser. La technique a été améliorée avec les Fast-R-CNN et les Faster R-CNN.

RNN : Recurrent Neuronal Networks ou réseaux de neurones récurrents. Ce sont des réseaux de neurones adaptés à l'analyse de signaux temporels comme la voix, du texte, un électro-cardiogramme ou le bruit d'une machine.

RPA : Robotic Process Automation, se dit des outils et méthodes d'automatisation des processus dans une entreprise consistant à piloter et enchaîner des applications existantes. Exploite de l'IA à des degrés divers selon le niveau de sophistication de l'automatisation, de ses fonctions d'apprentissage, et de traitement des images et du langage. Exemple classique : filtrage de mails entrants dans un service client.

Sciences cognitives : disciplines scientifiques dédiées à la description, l'explication et la simulation des mécanismes de la pensée humaine, animale ou artificielle. Les progrès dans ces domaines permettent d'améliorer les techniques utilisées dans l'IA.

Seq2seq : sequence to sequence, technique utilisée dans le traitement du langage dans les réseaux de neurones LSTM.

SGD : stockastic gradient descent, technique utilisée dans les réseaux de neurones pour déterminer le poids optimal des synapes.

Sigmoïde : fonction de normalisation de la valeur de sortie d'un neurone, la préservant entre 0 et 1.

Singularité de l'IA : moment symbolique où l'IA dépassera le niveau d'intelligence humaine. Mais est-ce que cela sera un moment précis ou un continuum ?

Softmax : autre fonction de normalisation de valeur de sortie d'un neurone dans un réseau de neurones. Présente l'intérêt de correspondre à une distribution statistique de probabilités pour un ensemble de neurones d'une couche données d'un réseau de neurones.

Sparsity : parcimonie, caractérise les matrices creuses, voir ce terme.

Supervisé : type d'aprentissage de machine learning et de deep learning qui exploite des données labellisées. Par exemples, des photos avec le nom (label) des objets qu'elles contiennent.

SVM : support vector machines, technique de segmentation utilisée dans le machine learning.

Symbolisme : méthodes et techniques de l'IA visant à représenter l'information et la savoir par des concepts organisés hiérarchiquement et par relations fonctionnelles et à haut niveau.

Synapses : liaisons entre neurones au niveau de la liaison entre axones et dendrites.

Synaptique : autre appellation des processeurs neuromorphiques.

Systèmes experts : systèmes d'IA s'appuyant sur la modélisation du savoir à haut niveau avec des logiques de prédicat (si ceci alors cela, ceci est dans cela, …) et des moteurs de règles.

TPU : Tensor Processor Unit, les processeurs neuromorphiques de Google, utilisés dans leurs data centers et aussi par DeepMind pour AlphaGo.

Transformers : technique de plus en plus utilisée dans les réseaux à mémoire pour le traitement du langage, capables de bien gérer le contexte dans lequel les mots sont interprétés et utilisant un mécanisme de gestion de l'attention. Ils exploitent des méthodes d'apprentissage non supervisé et analysent le langage de manière parallélisée et non par séquences.

Transhumanisme : courant de pensée ambitionnant de fusionner l'homme et la machine pour lui permettre de

dépasser ses capacités intellectuelles et d'atteindre l'immortalité.

TrueNorth : processeurs neuromorphique d'IBM.

Vie artificielle : simulation de la vie à un niveau d'abstraction arbitraire, via des logiciels.

VUI : Vocal UserInterface, l'interface vocale d'un agent conversationnel audio. Cela comprend l'ensemble des interactions avec l'utilisateur et leur qualité.

Watson : au départ, ordinateur d'IBM ayant gagné au jeu Jeopardy en 2011. C'est devenu depuis une plateforme logicielle d'IA appliquée à différents métiers et besoins. (chatbot, reconnaissance d'images, etc) qui sont notamment disponibles en cloud. C'est en pratique une grosse boîte à outils d'IA pour le traitement des images et du langage.

Historique des révisions du document

Version	Date	Modifications
1.0 (362 pages)	19 octobre 2017	Seconde version publiée sur http://www.oezratty.net.
1.01	24 octobre 2017	Ajout de Neuron Data dans l'historique des systèmes experts. Mise à jour du tableau de startups marketing de l'IA provenant de Fred Cavazza.
1.02	30 octobre 2017	Corrections orthographiques diverses.
1.03	7 novembre 2017	Ajout d'In Principio dans les médias spécialisés dans l'IA.
1.04	16 novembre 2017	Remplacement de Caffee par Caffe2 dans le tableau des outils de développement. Ajout de PyTorch.
2.0 (520 pages)	15 novembre 2018	Nouvelle édition 2018 entièrement refondue et réactualisée avec 13 mois d'actualités de l'IA.
2.1 (522 pages)	16 novembre 2018	Corrections au sujet des parts de marché Amazon Echo / Google Home, au sujet d'Antvoice, et de Search'XPR. Ajout d'Ezako, du projet de trains autonomes de la SNCF, Geo4cast et Score4Biz.
2.2	17 novembre 2018	Un peu de spellcheck et modifications au sujet de Scortex et de RefundMyTicket.
2.3	20 novembre 2018	Corrections au sujet de SANEF et Vinci.
2.4	22 novembre 2018	Compléments au sujet de Dhatim et de l'expertise comptable. Ajout d'un lien sur l'histoire des IA connexionniste et symbolique.
2.5	23 novembre 2018	Ajout de Habana et SambaNova dans les chipsets serveurs.
2.6	3 décembre 2018	Corrections sur « case law » vs « civil law », et sur Deeper Blue / Deep Blue.
3.0 (624 pages)	18 novembre 2019	Quatrième édition 2019. Révision de toutes les parties avec 12 mois d'actualité. Notamment celle sur l'aviation, sur la distribution, sur l'éthique de l'IA. Plus de détails sur les R-CNN, sur les biais des données. Mises à jour sur les processeurs de l'IA. Ajout d'une partie sur Graph Neural Networks, les AIOps, les achats, sur l'art, sur la recherche, sur les fake news, sur le travail collaboratif. Ajout des deep forests dans les méthodes de classification du machine learning. Ajout de l'Estonie, de Taïwan et Singapour. Ajout d'une petite bibliographie. Déplacement des hyperliens dans les notes de bas de page.

4.0 (742 pages)	Février 2021	Cinqième édition 2021. Ajout de Dynatrace dans les acteurs des AIops. De Ponycode dans les outils de développement. D'Expert System dans les outils conversationnels. D'Advalo dans le retail. Explication des notions d'epoch, batches et itérations dans la partie sur la rétropropagation des gradients. Ajout d'une rubrique sur le federated learning. Grande mise à jour sur les chipsets couvrant notamment l'offre de Nvidia, Graphcore, Intel, Xylinx et plein d'autres acteurs comme Enflame Technologies, Prefered Network, Pezy Computing, PQ Labs Inc, Untether AI et dans l'embarqué : NextVPU, Black Sesame, Perceive, Corerain Technologies, FuriosaAI, DinoPlusAI, XMOS, Yumain Sensing, Canaan et SiMaai. Ajout d'une nouvelle partie sur la big data et son lien avec l'IA. Ajout d'une partie sur l'empreinte énergétique de l'IA. Ajout d'une rubrique sur les applications de l'IA dans le domaine maritime. Ajout d'une rubrique sur les usages de l'IA autour du covid-19. Ajout d'une rubrique sur l'usage de l'IA pour développer des batteries. Ajout de Nucleai dans les startups en oncologie. Ajout de Hi! PARIS dans l'enseignement et la recherche en France. Glossaire : cybernétique, fonction de coût, matrices creuses, sparsity, transformers. Compléments dans la bibliographie.

Vous êtes lecteur, chercheur, expert, fournisseur et avez détecté des erreurs ou graves oublis dans ce document ? Il y en a sûrement ! N'hésitez alors pas à me contacter (olivier@oezratty.net) pour me les signaler. J'effectuerai alors des mises à jour de ce rapport tout en mettant à jour la chronologie dans le tableau *ci-dessus*.

Ce document est téléchargeable à partir de https://www.oezratty.net/wordpress/2021/usages-intelligence-artificielle-2021.