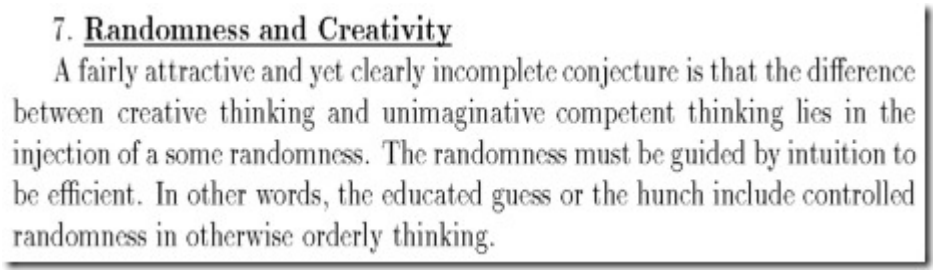
Opinions Libres
le blog d'Olivier Ezratty

Où en est l'IA émotionnelle ?

Dans le **fameux mémo de John McCarthy d'aout 1955**, les pères fondateurs de la discipline de l'intelligence artificielle en définirent les principes de base. Il s'agissait de transposer dans des machines une bonne partie des capacités d'intelligence humaine, notamment la compréhension du langage, la vision et le raisonnement.

L'émotion n'était pas encore au programme. Il comprenait tout juste l'intégration d'une dose d'aléatoire afin de générer de la créativité, une idée de Nathaniel Rochester d'IBM, l'un des quatre protagonistes à l'origine du Summer Camp de Darmouth.



7. **Randomness and Creativity**

A fairly attractive and yet clearly incomplete conjecture is that the difference between creative thinking and unimaginative competent thinking lies in the injection of a some randomness. The randomness must be guided by intuition to be efficient. In other words, the educated guess or the hunch include controlled randomness in otherwise orderly thinking.

L'IA se définissait comme *"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"*. L'IA relevait donc de la création de machines qui pensent, mais sans émotions qui sont le propre de l'homme. Au fil du temps, c'était d'ailleurs perçu comme un avantage par rapport à l'Homme. L'absence d'émotion implique une rationalité dont il ne fait pas toujours preuve dans ses processus de décision.

Faisons un détour dans cette intéressante Histoire du Summer Camp de Darmouth, via **Ray Solomonoff and the Dartmouth Summer Research Project in Artificial Intelligence**. Elle fut rédigée en 1996 par Grace Solomonoff, la femme – elle-même scientifique – de Ray Solomonoff, l'un des participants les plus actifs de cette épopée qui s'était tenue pendant huit semaines de l'été 1956 et qui avait réuni de 20 à 32 participants, certains n'étant que de passage.

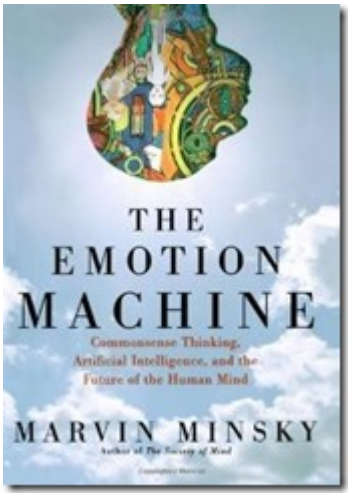
D'après les notes de Ray Solomonoff, ce hackathon de l'IA avait aboutit à quelques publications, des rencontres de chercheurs complémentaires dans cette nouvelle discipline, l'émergence de nouvelles idées, et celle selon laquelle elles pourraient aboutir dans un premier temps à la création de machines capables de résoudre des problèmes très spécialisés.

Deux courants cohabitaient et continuent de cohabiter dans l'IA : celui du symbolisme ou du raisonnement déductif basé sur la connaissance existante, que l'on retrouve dans les moteurs de règles et le systèmes experts, et celui du connexionnisme, ou du raisonnement inductif, qui tire

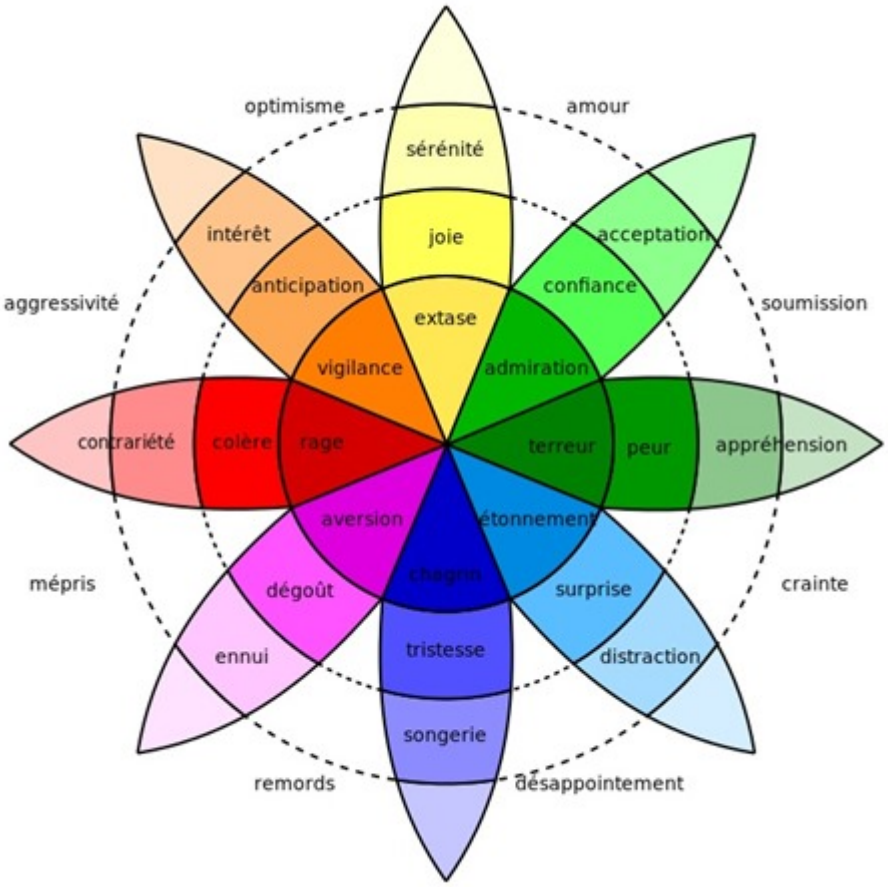
ses conclusions de l'observation des évènements, et que l'on retrouve aujourd'hui dans la discipline du machine learning et du deep learning. En gros, soit on connait déjà les règles de fonctionnement de tel système et on les utilise avec des moteurs de règles, soit on déduit ces règles de l'analyse de données pour faire des prédictions.

On en apprend un peu plus dans **Machines Who Think** de Pamela McCorduck (2003, 584 pages), une Histoire des premières décennies de l'IA. L'un des arguments contre la possibilité de créer des machines pensantes était qu'il était impossible qu'elles aient des émotions car celles-ci son éminemment biologiques. La robotique, surtout humanoïde, demandait cependant une certaine prise en compte d'émotions, intégrées dans leurs capacités d'intégration et de communication avec leurs utilisateurs humains. Au départ, cela relève d'un simple mimétisme. Cela peut se sophistiquer avec la capacité d'un robot d'engager une discussion pour entrer en communication. L'intégration sociale des robots est devenue une question en soi intégrant la notion d'intelligence émotionnelle.

Dans **The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind**, Marvin Minsky considérait que les émotions sont des attributs de l'intelligence qui la mettent en relief. Les émotions sont le résultat de l'activation de règles de fonctionnement par le cerveau, surtout limbique, en réaction à des évènements intérieurs ou extérieurs. La plus connue est le stress en réaction à un danger perçu. Il peut déclencher des réflexes de fuite ou de défense, avec des nuances qui dépendent des circonstances, donc de règles. Minksy mettait bien en évidence les relations complexes entre les émotions, les besoins et la pensée et la sophistication des processus d'apprentissage humains liés en grande partie à notre interaction avec notre environnement et avec d'autres humains.



Fast-forward en 2018. 62 ans après le Summer Camp de Darmouth, où en sommes-nous dans l'utilisation des émotions dans le cadre des usages de l'intelligence artificielle ? Les spécialistes ont une dénomination pour décrire l'IA émotionnelle : l'informatique affective ou Affective Computing. La segmentation des émotions humaine est d'ailleurs un sujet pas définitivement tranché. Il y en a de nombreuses, l'une des plus connues étant la roue de Plutchik (*ci-dessous*). Elle comprend huit émotions de base structurées en paires d'opposés : joie et tristesse, peur et colère, dégoût et confiance, et enfin, surprise et anticipation, avec trois niveaux d'intensité et des combinaisons par paires. Les modèles de cartographie d'émotion les plus simples se contentent de n'en conserver que quatre principale : la peur, la colère, la joie et la tristesse.



Nous n'avons toujours pas su créer de machines pensantes. Aucune IA ne sait raisonner de manière complexe, notamment par analogie. L'IA actuelle est dite étroite car elle est faite de solutions à des problèmes relativement simples et monolithiques. Nous sommes dans le domaine de l'intelligence augmentée et en (petites) pièces détachées. Le raisonnement sophistiqué est la première pierre d'achoppement de l'IA.

Il en va de même pour l'intégration des émotions dans l'IA. Celle-ci est de plus en plus généralisée, mais est tout autant mise en œuvre en pièces détachées qui fonctionnent plus ou moins bien selon les cas. La grande intégration n'est pas encore là. Pourtant, grâce à l'IA, les machines peuvent faire plein de choses avec nos émotions. Elles peuvent en détecter certaines, les interpréter, parfois les expliquer, réagir face aux émotions humaines, en afficher dans le cas des outils qui interagissent avec les humains, et même interagir avec l'Homme pour déclencher chez lui des émotions et le faire agir en conséquence. C'est déjà fort impressionnant ! Au bout du compte, ces différentes techniques servent notamment à créer des chatbots capables de conduire des discussions réalistes avec leurs utilisateurs et aux robots d'avoir une forme élémentaire de sociabilité.

L'objet de ce texte est de faire un voyage rapide dans le lien entre l'IA et nos émotions puis de conclure avec quelques-unes des nombreuses questions éthiques posées par tous ces outils. Pour chaque exemple, je m'appuyer sur des travaux de recherche et/ou des solutions du commerce issues de grandes entreprises ou de startups.

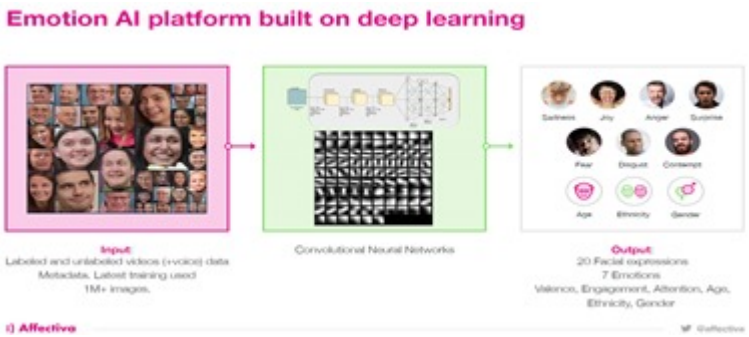
IA qui détecte les émotions humaines

La détection des émotions humaines est un vaste champ assez mature de l'IA. Il s'appuie sur divers capteurs : vidéo, micros et biométriques qui permettent d'en savoir beaucoup sur nous. Au départ, les capteurs et logiciels dédiés à la détection des émotions fonctionnaient aussi en pièces

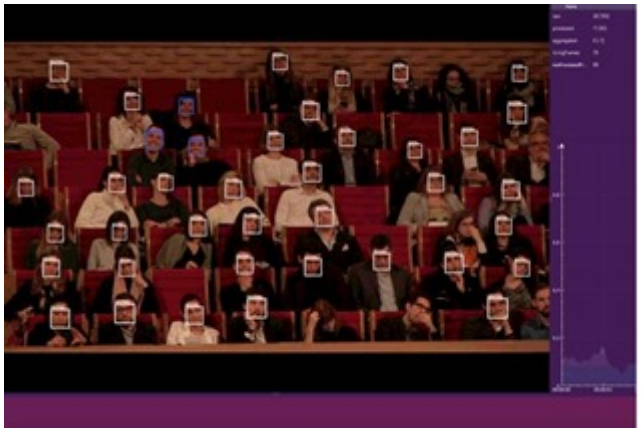
détachées. L'heure de l'intégration est en train d'arriver.

La reconnaissance des émotions dans le visage à partir de caméras est assez ancienne. Elle est standardisée par le système de description FACS pour **Facial Action Coding System**, créé en 1978 par les psychologues américains Paul Ekman et Wallace Friesenen.

J'avais découvert la startup Américaine **Affectiva** au CES 2013 (voir le **Rapport CES 2013**, page 256). Cette startup présentait une solution de captation les émotions d'un utilisateur exploitant une simple webcam sur un micro-ordinateur. Il valorise un projet de recherche du MIT Media Lab et a depuis levé $26,3M. La startup visait le marché de la publicité et du retail mais a eu du mal à le pénétrer. Depuis, ils s'intéressent au marché de l'automobile pour détecter l'état du conducteur, comme le manque d'attention, qui est une fonctionnalité finalement assez limitée. Ils annoncent qu'ils ont entraîné leurs modèles de deep learning avec 5 millions de visages issus de 75 pays. Leur logiciel évalue les paramètres suivants : joie, gaieté, dégoût, mépris, peur, surprise, colère ainsi que la valence (allant du négatif au positif), l'engagement et l'attention, reprenant la structure de la roue de Plutchik. Le tout exploite l'analyse de 20 expressions faciales différentes via des réseaux convolutifs (ou convolutionnels) CNN et des SVM, l'une des plus courantes méthodes de segmentation du machine learning. Mais il est difficile de savoir où cette solution est déployée d'un point de vue pratique.



En France, la société **Datakalab** fait encore mieux qu'Affectiva en analysant 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 ! 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.



La société française **XXII** présentait au CES 2018 sa plateforme d'intelligence artificielle *bio-inspirée* destinée au retail, à la sécurité et aux véhicules autonomes. Elle exploite des algorithmes de reconnaissance d'émotions et de micro-expressions, de reconnaissance et identifications de produits, de suivi de personnes et de reconnaissance et identification de gestes et de comportements pour identifier les agressions, chutes et autres dangers (**vidéo**).

L'Américain **Hire*View** (2004, $93M) exploite les vidéos d'interviews en ligne de candidats au recrutement. Sa solution analyse leurs visages et identifie des traits de personnalité. La solution est déployée chez Unilever aux USA, à une échelle difficile à apprécier. On arrive aux frontières acceptables de l'éthique avec ce genre de solution ! C'est au propre comme au figuré du recrutement à la tête du client ! Mais la solution pourrait aussi être utilisée pour s'entraîner, avec une boucle de feedback.

Et ce ne sont que quelques exemples !

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, voir l'étude européenne **Survey on Emotional Body Gesture Recognition**, publiée en janvier 2018 (19 pages) et qui fait un état des lieux. 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.

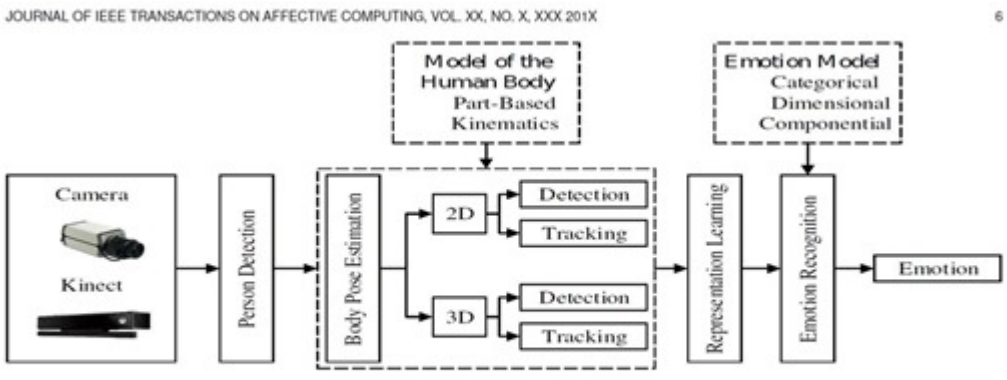


Figure 7. General overview of an Emotion Body Gesture Recognition system. After detecting persons in the input for background extraction, a common step is to estimate the body pose. This is done either by detecting and tracking different parts of the body (hands, head, torso, etc.) or by mapping a kinematic model (a skeleton) to the image. Based on the extracted model of the human body, a relevant representation is extracted or learned in order to map the input to a predefined emotion model using automatic pattern recognition methods.

Encore plus surprenants, ces travaux de recherche publiés dans **DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition using Low-Cost Thermal**

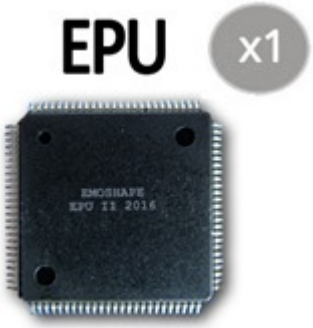
Imaging in Unconstrained Settings en aout 2017, visent à détecter le niveau de stress d'utilisateur avec des caméras infrarouges analysant le rythme de respiration des humains avoisinants. Le tout exploite un simple réseau de neurones convolutif !

Après les visages et les gestes, on commence à exploiter la voix pour détecter des émotions, au-delà des fonctions habituelles de reconnaissance de la parole. La startup israélienne **Beyond Verbal** est spécialisée dans la détection des émotions dans la voix. Elle l'exploite aussi pour identifier des pathologies neurodégénératives émergentes chez les personnes âgées. La startup parisienne **Batvoice** utilise aussi la voix pour détecter l'état émotionnel de clients et d'opérateurs dans les centres d'appels. C'est aussi le cas de la startup israélienne **Nemesysco** qui utilise sa solution d'analyse des émotions dans la voix dans divers marchés y compris pour sonder les collaborateurs d'une entreprise et détecter d'éventuels risques en termes de sécurité ou de fraudes. C'est en quelque sorte un détecteur de mensonges. La fraude est également détectée dans les appels de déclaration d'incidents aux assurances. Mais on manque de données indépendantes validant ou invalidant ces différentes solutions.

Une équipe chinoise vient de publier récemment **A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks** et propose une méthode permettant d'analyser plus efficacement les émotions dans la voix… en chinois, histoire d'éviter de générer un trop gros taux d'erreurs !

Comme nous l'avons vu avec Datakalab, la biométrie sert aussi à détecter les émotions, qu'il s'agisse de montres connectées mesurant le rythme cardiaque ou la transpiration ou des casques de captation d'ondes électroencéphalogrammes. Une équipe de recherche anglo-malésienne publiait ainsi **Emotion-Recognition Using Smart Watch Accelerometer Data: Preliminary Findings** fin 2017 et illustrait comment capter des émotions avec l'accéléromètre d'une montre connectée. Au CES 2018, j'avais repéré les Japonais **Imec** et le **Holst Centre** qui démontraient un casque de captation EEG capable de détecter les émotions des utilisateurs. Le casque s'appuie sur un algorithme de machine learning développé par l'Université d'Osaka. C'est devenu une pratique assez courante, mais à des fins expérimentales ou pour des applications spécifiques. Et pour cause, un utilisateur ne peut pas trimballer toute la journée un casque EEG sur sa tête ! Par contre, la nuit, pourquoi pas ! C'est la technique utilisée par la startup française **Rythm** avec son casque **Dreem** qui doit vous aider à mieux vous endormir.

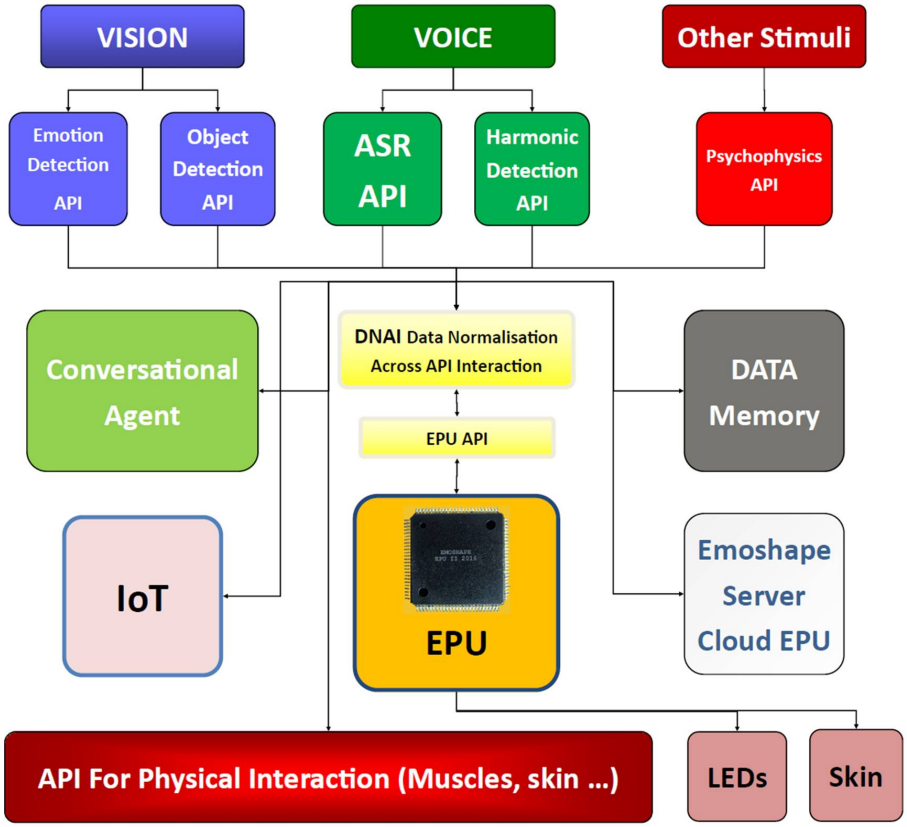
Enfin, on peut aussi capter les émotions d'utilisateurs en analysant leur production écrite. C'est ce que font un paquet de startups qui analysent les sentiments dans les réseaux sociaux ou la qualité des CV. Les startups françaises **Natural Talk** (2016) et **Cognitive Matchbox** (2016) proposent chacune une solution de routage d'appels optimisée aux centres d'appels qui analyse la personnalité et les émotions des clients via leurs messages écrits afin de les orienter vers le meilleur agent. Elles exploitent les fonctions 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.



La détection des émotions passe sinon de plus en plus par l'intégration des données issues de plusieurs capteurs (vidéo, audio, autres). On appelle cela la captation multimodale d'émotions. Elle est bien décrite dans **Affective computing From unimodal analysis to multimodal fusion**, publié par une équipe de recherche anglo-singapourienne en février 2017 (28 pages).

Cela permet d'affiner les résultats et d'obtenir des indicateurs plus précis. Par contre, cela nécessite des jeux de données d'entrainement conséquents que les startups du secteur n'ont pas toujours à leur disposition.

Dans ce registre, la startup New-Yorkaise **Emoshape**, créée par le Français Patrick Levy-Rosenthal, présentait au CES 2018 son composant électronique Emotion Processing Unit (EPU, *ci-dessus*), destiné à déterminer en temps réel les émotions des utilisateurs et à permettre aux robots et autres applications de répondre avec un état émotionnel en phase avec celui de l'utilisateur (**explication**). Le chipset va récupérer 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 permettre 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. 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 vidéos, 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. Histoire par exemple de ne pas vous déprimer plus si vous l'êtes déjà !

Bref, pour ce qui est de l'expression extérieure de nos émotions, peu de choses semblent pouvoir échapper aux capteurs, tout du moins en théorie ! Reste à savoir ce que les IA peuvent en faire ! Evidemment, tous ces capteurs et outils d'interprétation ne lisent pas dans nos pensées, là où sont logées nos réelles émotions. Nos expressions les trahissent parfois, mais pas systématiquement et pas forcément avec suffisamment de précision.

IA qui interprète et comprend les émotions humaines

Une fois que l'on a détecté les émotions extérieures avec un ou plusieurs des différents capteurs et systèmes évoqués ci-dessus, il faut les interpréter. Les émotions ont du sens en fonction du contexte. Elles dépendent aussi de la culture et de la langue, pour ce qui est de la voix et de son intonation tout comme de la gestuelle.

Des outils à base d'IA peuvent analyser la corrélation entre les émotions et les évènements qui les génèrent. Cela permet par exemple d'évaluer l'impact de contenus, dans la publicité ou dans la fiction. Ces techniques reposent le plus souvent sur du machine learning. Pour le cas les plus complexes, comme dans l'entraînement dit non supervisé de chatbots, l'entrainement peut s'appuyer sur des réseaux de neurones. Ils vont permettre d'ajuster la nature des réponses aux questions en fonction du contexte émotionnel du dialogue entre le chatbot et l'utilisateur.

L'interprétation intervient également dans l'évaluation des émotions générées par des contenus comme de la musique ou toute autre forme de création. Dans le cas de créations générées par des outils à base d'IA, cela permet de créer une boucle de feedback entre création à base d'IA et

utilisateurs, histoire de déterminer les contenus générés qui ont le meilleur quotient émotionnel ! J'avais évoqué cette possibilité dans un précédent article, **L'IA est-elle vraiment créative ?** de novembre 2017. Ces méthodes pourraient aussi servir à évaluer la pertinence de formes d'humour générées par des IA, essentiellement des agents conversationnels. L'humour repose souvent sur l'exploitation d'analogies. Certaines fonctionnent bien et d'autres non. Ce qu'une boucle de feedback permet d'évaluer avant qu'une IA soit capable es-abstracto d'exploiter des recettes miracles et répétables de l'humour.

IA qui agit en fonction des émotions humaines

Une fois détectées et interprétées, les émotions captées doivent servir à quelque chose ! Qu'il s'agisse d'un véhicule autonome, d'un robot ou d'un agent conversationnel, ces outils peuvent exploiter leur interprétation des émotions pour réagir. La prise en compte des émotions peut aussi intervenir dans les jeux vidéos. Le "gameplay" peut s'appuyer sur des moteurs de règles qui sont intégrés dans le scénario des jeux. C'est devenu d'ailleurs une discipline à part entière de l'IA.

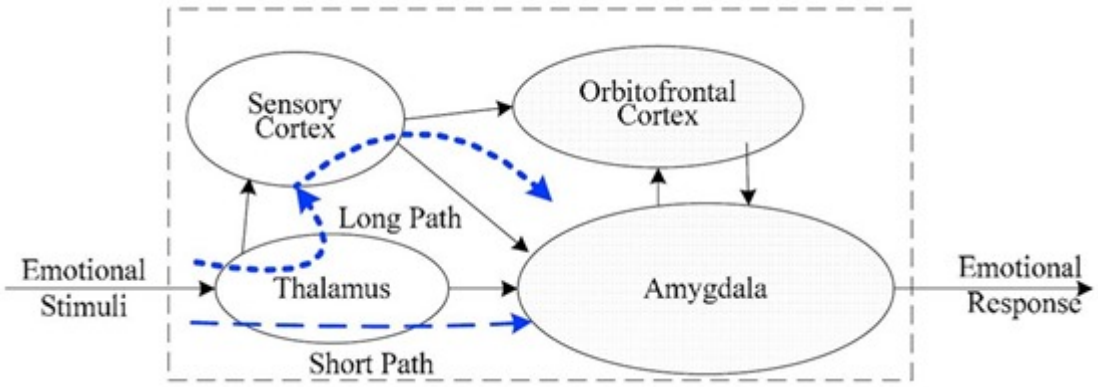
L'adaptation des modes d'interaction des agents conversationnels aux émotions dégagées par les utilisateurs est dans le domaine possible. Dans un cas simple, un agent vocal peut adapter son débit de parole à celui de l'utilisateur. S'il sent que l'utilisateur est pressé, il peut générer des réponses plus succinctes que si l'utilisateur semble avoir du temps. Lorsque vous arrivez au travail, l'IA de votre poste de travail exploitant sa webcam pourrait détecter votre humeur et vous distraire pour vous mettre de bon poil.

Les actions d'un agent qui interagit avec un utilisateur peuvent aussi exploiter une mémoire des émotions. L'agent n'aura pas forcément accès à toute l'histoire de la vie de l'utilisateur comme ses traumatismes d'enfance ou ses souvenirs. Il pourra par contre accumuler des souvenirs de ses interactions avec l'utilisateur. Ceux-ci permettent à l'agent d'adapter son comportement via de l'apprentissage. Peu de systèmes ont aujourd'hui ce genre de capacités mais rien n'empêche de les développer.

Des outils interactifs peuvent aussi nous aider à ajuster le niveau émotionnel de nos propres productions. Ainsi, l'outil **DeepBreadth de Google**, qui n'a rien à voir avec le DeepBreadth évoqué précédemment, conseille les utilisateurs sur l'attitude à adopter en écrivant des réponses à des emails. Il prévient l'utilisateur d'un niveau d'agressivité inapproprié ! Ceci date d'avril 2017 et prend la forme d'un plugin de Chrome ! Il est exploitable sous forme d'APIs en cloud par des développeurs d'applications !

Nous sommes ici face à des scénarios potentiels extrêmement divers où la technique croise l'éthique. D'un côté, nous avons souvent besoin de machines qui calculent et raisonnent sans émotions, et de manière uniquement rationnelle, même si des paramètres de décision de nature émotionnelle peuvent leur être injectées. De l'autre, des systèmes sophistiqués peuvent être créés qui utilisent les émotions humaines pour les manipuler à des fins plus ou moins acceptables. Cela peut aller de tout ce qui permet de vendre quelque chose, de modifier l'image d'une marque, jusqu'à influencer nos choix citoyens et politiques. Les méthodes Russes employées à l'occasion du Brexit ou de la présidentielle US de 2016 ne sont qu'un avant-gout de ce qu'il serait possible de faire de manière très sophistiquée dans ces registres.

A contrario, dans le moins contestable, des méthodes de psychothérapies curatives à base d'IA sont parfaitement envisageables. Un peu dans la lignée du scénario du film Her.



La science pas si fiction que cela peut aller encore plus loin. On sait par exemple que diverses sociétés telles que Neuralink, créée par Elon Musk, travaillent sur la connexion entre le cerveau humain et l'IA. J'avais eu l'occasion de creuser le sujet dans une série de trois articles sur "**Ces startups qui veulent bidouiller le cerveau**" en mai 2017. J'étais particulièrement sceptique sur la capacité de ces projets de modifier notre mémoire dans le cortex préfrontal, du fait de l'extrême complexité de l'organisation des neurones et de leur nombre. Notre mémoire est individuelle et distribuée dans l'ensemble du cortex préfrontal, visuel, auditif et moteur dans des milliards de neurones et des centaines de milliards de connexions neuronales (synapses/dendrites).

Par contre, les techniques à base d'électrodes sur lesquelles planche **Neuralink** pourraient très bien servir à altérer le fonctionner de parties du cerveau limbique qui génèrent la sécrétion d'hormones. Par exemple, pour nous rendre heureux ou malheureux, stressé ou calme. La dopamine est produite dans le mésencéphale à la base du cerveau, l'ocytocine vient de l'hypothalamus, l'adrénaline est produite par les glandes surrénales qui sont activées nerveusement par le cerveau, l'endorphine est générée par l'hypophyse dans le cerveau, le cortisol est produit dans les glandes surrénales par activation via l'hormone ACTH générée dans d'hypophyse et la mélatonine est générée dans l'épiphyse dans le cerveau limbique. C'est un peu plus plausible que l'écriture dans le cortex car il s'agit d'activer des zones neuronales relativement simples, des sortes de robinets. Ces zones du cerveau limbiques sont par contre logées au centre du cerveau et plus difficiles d'accès.

Ces techniques ont un double versant : elles pourraient servir à traiter des pathologies comme le PTSD (Post Traumatic Stress Disorder) qui affecte par exemples les militaires revenant de théâtres d'opération difficiles. Mais elles pourraient aussi servir à contrôler des individus pour les conditionner ou leur faire commettre des actes moralement répréhensibles, et même influencer leur vote lors d'élections ! Bref, danger !

IA qui affiche des émotions

Avant-dernier sujet de ce long inventaire des capacités émotionnelles de l'IA, celui de l'affichage d'émotions par des IA, notamment via des agents conversationnels ou des robots. Les afficher revient d'abord à les simuler, pas forcément à en avoir. Ce sont des moyens d'anthropomorphiser des interactions avec les utilisateurs en utilisant les codes émotionnels de ces derniers. Mais l'étude des animaux, notamment domestiques, montre que les animaux sont capables d'émettre des émotions riches sans maitriser le langage. C'est une source d'inspiration qui permet aux chercheurs d'élaborer des méthodes d'expression corporelle pour les robots. Comme l'indique Jean-Marc Fellous dans **From Human Emotions to Robot Emotions** (2004), la principale fonction de l'affichage d'émotions est de communiquer des informations de manière simplifiée et

efficace.

La parole synthétique devrait être un moyen d'émettre des émotions verbales mais on est encore très loin du compte. On peut le voir avec les solutions les plus avancées comme celle de **Lyrebird**, notamment dans cette **fameuse vidéo** mettant en scène un Barack Obama de synthèse, à la fois en vidéo et en audio. Qui plus est, l'ancien Président américain est plutôt "dans le contrôle" et dans la maitrise de ses émotions, donc l'effet émotionnel est faible dans ce genre de prouesse technique.

Dans **Emotional End-to-End Neural Speech synthesizer**, publié en novembre 2017, des chercheurs coréens utilisent des réseaux de neurones récurrents pour de générer une parole de synthèse capable d'émettre des émotions de manière plus réaliste, mais ces progrès sont très incrémentaux. D'autres chercheurs travaillent sur la génération de gestes artificiels accompagnant la voix (cf **Speech-Driven Animation with Meaningful Behaviors**, 2017).

.

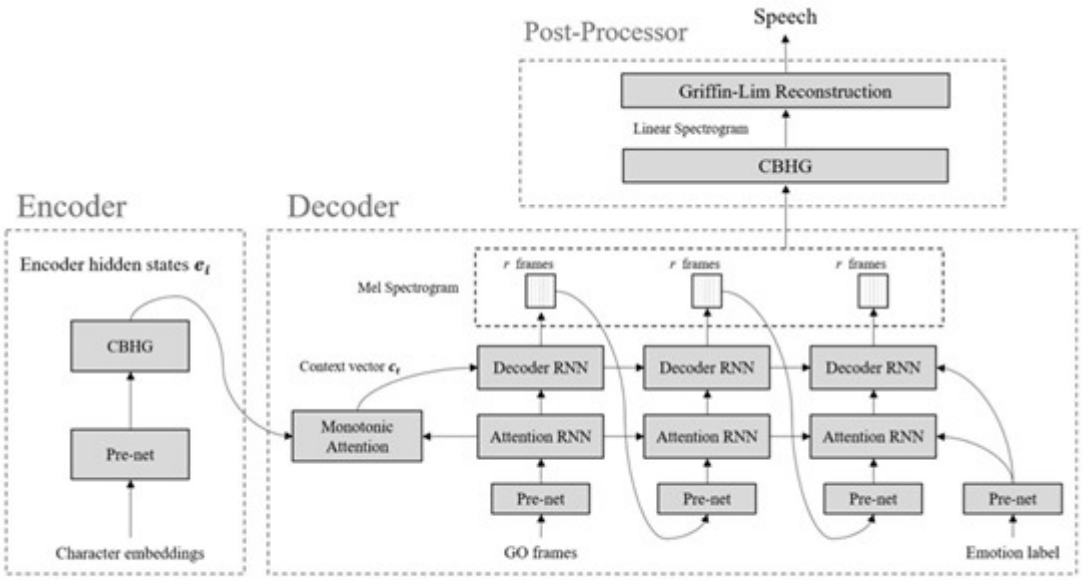
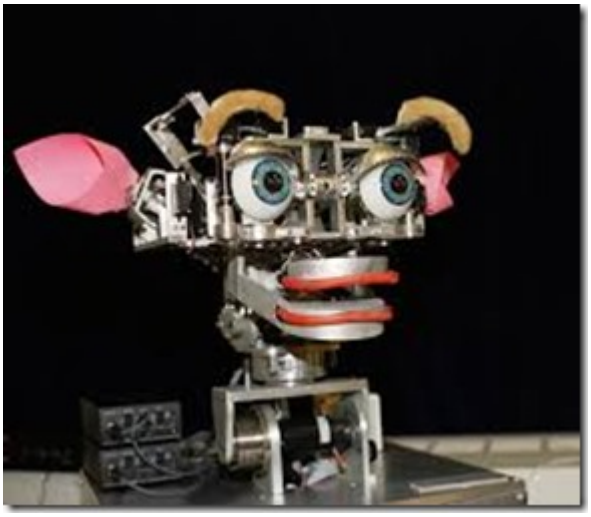


Figure 1: Emotional end-to-end speech synthesizer.

L'expression d'émotions passe aussi par les robots, notamment par la gestuelle comme avec celles de **Nao** d'Aldebaran Robotics / Softbank Robotics ou même celle des différentes générations de chiens **Aibo** de Sony. Les yeux doux du **Buddy** de Blue Frog Robotics, présentés sur un écran de tablette vont dans le même sens mais de manière assez rudimentaire, tout comme son équivalent un peu plus mobile chez **Spoon.ai**. Il en va de même du robot **Honda** 3E-A18 présenté au CES 2018. C'est un robot compagnon doué d'émotions dans les expressions du visage qui ne sont que des images simplistes provenant d'un écran intégré dans le haut de sa carcasse (*ci-dessous*).

On traverse la vallée de l'étrange (uncanny valley) avec les robots intégrant un visage animé, une pratique courante chez les roboticiens japonais ou avec le fameux robot **Sofia** de Hanson Robotics, une startup créée par David Hanson, un Américain, et installée à Hong Kong. Sofia a un visage qui essaye d'exprimer des émotions, mais cela reste rudimentaire, essentiellement pour des raisons mécaniques. Voir **l'interview de Sofia par Jimmy Fallon** qui date d'avril 2017 et **une autre**, qui date de janvier 2018 au CES 2018. Sans grand progrès apparent ! Sachant que les démonstrations de Sofia sont parait-ils toutes scriptées. Bref, de la fake IA !

Nous sommes en fait encore très loin de la vallée de l'étrange, ce syndrome du malaise que l'Homme pourrait ressentir en interagissant avec un hypothétique robot qui serait trop humain dans sa forme et son expression. D'où la création de robots qui n'ont pas du tout l'aspect humain, comme le fameux **Kismet** créé à la fin des années 1990 au MIT par Cynthia Breazeal (*ci-dessous*). Ce syndrome de la vallée de l'étrange pourrait cependant se manifester avec les formes les plus avancées de robots sexuels où la plastique et le réalisme corporel comptent plus que pour les robots conversationnels.



Il y aurait 43 muscles dans nos visages permettant d'exprimer jusqu'à 10 000 émotions différentes ! Les robots en sont encore loin. C'est d'abord un problème de mécanique, puis d'IA pour animer convenablement cette mécanique. D'où l'intérêt de travaux de recherche sur la

création de muscles artificiels, comme **ceux de Harvard et du MIT**.

Mais la production d'émotions synthétiques est tout à fait possible dans l'immatériel. Elle peut passer aussi bien par la création musicale que par la création de scénarios de fictions écrites ou pour l'audiovisuel, qui exploitent nos émotions. Leur conception doit cependant passer par une boucle de feedback humaine, ce qui relève donc de l'apprentissage supervisé.

IA ayant des émotions

La question suivante est de se demander si les IA peuvent avoir leurs propres émotions. Comme l'Homme Bicentenaire, joué par Robin Williams en 2000, qui l'un des rares films de science fiction de robots qui ne soit pas dystopique.

Les chercheurs en IA et en neurosciences s'accordent pour l'instant sur le fait que les formes actuelles d'IA ne peuvent pas en avoir. Elles n'ont pas de corps, pas d'hormones, pas de mémoire de leur interaction avec le monde et ne sont pas passées par le processus d'apprentissage de la vie. Elles n'ont pas de mémoire émotionnelle équivalente à celle de l'Homme avec sa construction qui démarre dans l'enfance et se poursuit avec l'apprentissage de la vie dans l'adolescence puis l'âge adulte. Elles ne détectent pas véritablement nos émotions. Elles ne font que détecter l'apparence externe de nos émotions.

Les robots n'ont pas d'envies, de désirs à part le besoin d'énergie pour fonctionner. Ils ne connaissent pas la perspective de la mort, sauf dans la science-fiction comme le **HAL** de 2001 Odyssée de l'Espace qui ne veut pas mourir (aka: être débranché).

Pourtant, comme nous venons de le voir, les IA peuvent interpréter mécaniquement nos émotions sans en avoir et interagir avec les Hommes en simulant de l'empathie. Cela reste encore une communication de forme asymétrique. En pratique, si on ne sait pas encore doter un programme de conscience ou lui permettant de ressentir des émotions, on peut toutefois les simuler. Mais aucun chatbot ne passe pour l'instant le test de Turing, consistant à pouvoir se faire passer pour un Humain.

Est-ce que les travaux de création d'AGI (intelligence artificielle générale) permettraient de lui adjoindre une capacité émotionnelle ? Ce n'est pas le sens des travaux qui semblent lancés. Les startups qui planchent sur l'AGI visent à créer des sortes de systèmes experts capables de résoudre des problèmes très complexes, mais en conservant un raisonnement rationnel. La

rationalité ultime est justement de ne faire preuve d'aucune d'émotion !

Ethique de l'IA émotionnelle

Ce petit tour de l'IA émotionnelle illustre un leitmotiv des usages potentiels de l'IA : comme toute nouvelle technologie, on y trouve le meilleur comme le pire. Certaines applications citées ci-dessus sont déjà très limites côté éthique, comme avec ces systèmes d'analyse de la personnalité qui s'attachent plus aux apparences visuelles voire auditives qu'à l'histoire des individus ou à leur intellect.

Les recherches sur l'éthique de l'IA et des robots sont nombreuses. En France, **Laurence Devillers**, chercheuse au CNRS-LIMSI, et auteure de "Des robots et des hommes", fait à la fois le tri entre les mythes, fantasmes et réalité de l'état de l'art en robotique et promeut une ligne de conduite éthique dans leur exploitation.

Dans **Ethical issues in affective computing**, l'Anglais **Roddy Cowie** fait aussi un tour assez complet de la question de l'éthique de l'IA émotionnelle.

En pratique, les réflexions sur l'éthique de l'IA sont fortement influencées par la science fiction. Comme ses auteurs ont tendance à privilégier les dystopies aux utopies, nous avons l'embarras du choix pour identifier des scénarios qui prêtent à la réflexion. C'est particulièrement visible dans la série **Blackmirror** qui imagine le pire dans tout un tas de situations d'un futur plus ou moins éloigné, intégrant généralement une association d'IA, de robotique et de mondes virtuels. Est-ce que l'imagination du pire permet de l'éviter ? L'histoire récente montre que cela dépend.

Le chemin le plus court vers la démocratisation des innovations technologiques associe les besoins humains et ceux des entreprises. Les fondements du capitalisme se souciant peu de morale, rien n'est bien sûr. Si la déshumanisation de la relation client via des chatbots permet de faire des économies, les entreprises l'adoptent tout de même. Elles ne mesurent pas forcément l'impact émotionnel de leurs choix sur les clients. Et si elles le font, c'est avec un temps de latence et il peut être alors difficile de faire marche arrière.

Cette démarche se retrouve dans les scénarios alambiqués construits par le collectif Utopia Dystopia de Nantes, que vous pouvez parcourir dans le compte-rendu **Interrogeons les futurs de l'intelligence artificielle** (67 slides). Il contient divers scénarios de manipulations de nos émotions à base d'IA et les réactions éthiques qu'ils peuvent générer.

Les chercheurs en robotique recommandent d'intégrer la dimension éthique dans leurs propres travaux. C'était la recommandation de la CERNA (Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique créée fin 2012 par l'alliance Allistene des sciences et technologies du numérique associant notamment le CEA, le CNRS, l'Inria et l'Institut Mines-Télécom) dans **Ethique de la Recherche en Robotique** (2014) qui "*préconise que les établissements ou institutions de recherche se dotent de comités d'éthique en sciences et technologies du numérique, traitant au cas par cas les questions opérationnelles[…]. Le chercheur doit prémunir les systèmes qu'il conçoit contre les effets indésirables, cela prévaut d'autant plus que les robots sont dotés d'une autonomie croissance. La confiance que l'on peut placer dans un robot, les possibilités et limites de celui-ci et du couple qu'il forme avec l'utilisateur, la reprise en main, le traçage – c'est-à-dire la possibilité de rendre compte du comportement – sont à considérer du point de vue éthique dans la conception du robot. Par l'imitation du vivant et l'interaction affective, le robot peut brouiller les frontières avec l'humain et jouer sur l'émotion de manière inédite. Au-delà de la prouesse technologique, la question de l'utilité d'une telle ressemblance doit se poser, et l'évaluation interdisciplinaire de ses effets doit être menée, d'autant plus que ces robots seraient placés auprès d'enfants ou de personnes fragiles*".

Ceci va dans le bon sens. Il faudra cependant éviter de focaliser ces questions d'éthique sur les robots car elles sont également à prendre en compte dans nombre d'usages qui n'y font pas appel. Qui plus est, on peut très bien avoir une recherche parfaitement éthique mais une innovation qui ne l'est pas. L'assemblage des technologies dans les usages est plus le fait des entreprises et des startups que des chercheurs. Les effets débilitants de certains réseaux sociaux ne proviennent pas de travaux de chercheurs et plutôt de méthodes de *growth hacking* jouant avec notre dopamine !

Et la régulation ? Dans un premier temps, le droit générique est naturellement applicable. Sont condamnables les abus de faiblesse, les solutions ayant pour objet de manipuler les utilisateurs, pour les faire réaliser des actions à l'insu de leur plein gré. La loi s'intéresse cependant peu à l'éthique et à la morale. Reste à affiner cela pour ce qui est de la manipulation des émotions utilisée comme arme politique ou arme de guerre froide ! Comme toutes les technologies bivalentes, le législateur devra à la fois préserver la capacité d'innovation de l'écosystème entrepreneurial tout en prémunissant les citoyens contre des effets les plus délétères. La frontière entre les deux est des plus ténue et cela risque de perdurer !

Cet article a été publié le 2 mars 2018 et édité en PDF le 13 mai 2019.