Opinions Libres

le blog d'Olivier Ezratty

# L’impact de l’IA sur la cybersécurité

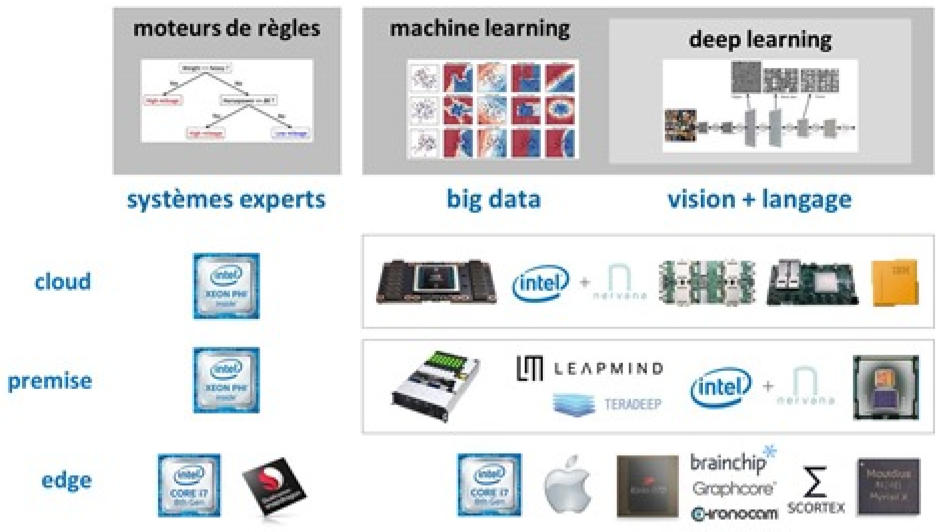
J’intervenais en ouverture du forum annuel du Cert-IST mercredi 29 novembre 2017 à Paris pour dresser un tableau général de l’impact de l’intelligence artificielle sur la cybersécurité (**présentations**).

Le **Cert-IST** – *Computer Emergency Response Team – Industrie, Services et Tertiaire* – est une association qui émane de Thales et qui partage les bonnes pratiques dans la cybersécurité entre ses entreprises françaises membres. On y trouve des banques, opérateurs télécoms, des services publics et divers industriels, notamment des secteurs de l’aérospatial et de la défense. Mon intervention avait été proposée par une autre association, le **CLUSIF** – Club de la Sécurité de l’Information Français – qui associe des offreurs de solutions de cybersécurité et des utilisateurs de ces solutions. C’est un peu l’équivalent d’un mixte Syntec Numérique + CIGREF (club des DSI de grandes entreprises) appliqué à la cybersécurité.

Mon intervention et celles auxquelles j’ai pu assister lors de ce forum me donnent donc l’occasion de faire un point rapide sur ce sujet épineux de l’impact de l’IA sur la cybersécurité. Le forum mettait très bien en évidence des aspects très pratiques et concrets des usages ou dérives de l’IA, bien loin des thèses fumeuses de la singularité.

Dans mon intervention, j’ai commencé, comme d’habitude, par segmenter les technologies de l’IA en grandes composantes : les moteurs de règles, le machine learning, le machine learning à base de réseaux de neurones, puis le deep learning qui utilise des réseaux de neurones profonds, et enfin, les agents et réseaux d’agents qui permettent d’assembler les briques de l’IA pour créer des solutions : robots, véhicules autonomes, chatbots et autres.

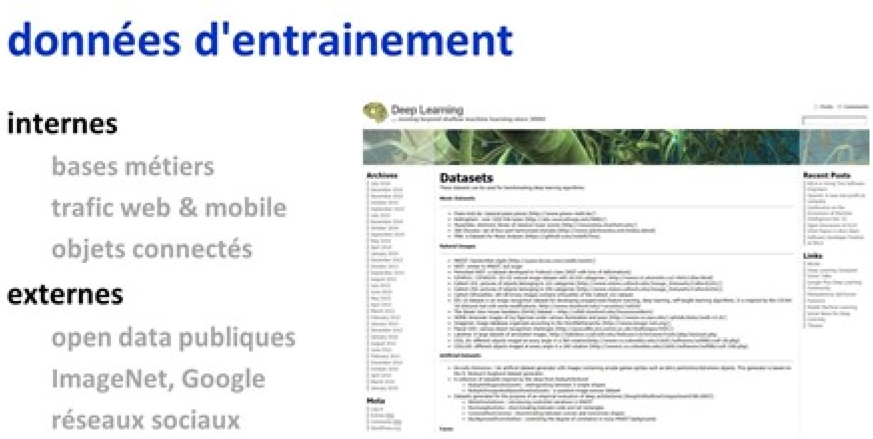
J’ai surtout expliqué comment fonctionnaient les réseaux de neurones profonds avec le cas de la reconnaissance d’images, puis expliqué comment ces réseaux de neurones étaient mis en œuvre côté matériel dans des processeurs spécialisés (GPU et processeurs neuromorphiques) qui font leur apparition à la fois du côté des serveurs, notamment pour les phases d’entrainement des réseaux de neurones, et du côté des objets connectés et mobiles, pour leur exécution. Avec les nouveaux risques que cela peut générer. Il y a d’ailleurs un lien étroit, dans la cybersécurité, entre l’IA, les objets connectés en tout genre et les réseaux télécoms. Ils constituent un continuum qu’il faut sécuriser de bout en bout ! Et les vulnérabilités à traiter sont très nombreuses à chacune des étapes de la chaine qui alimente les solutions exploitant des techniques logicielles à base d’IA. Cela affecte surtout celles qui fonctionnent avec des modèles probabilistes comme dans tout le périmètre du machine learning.



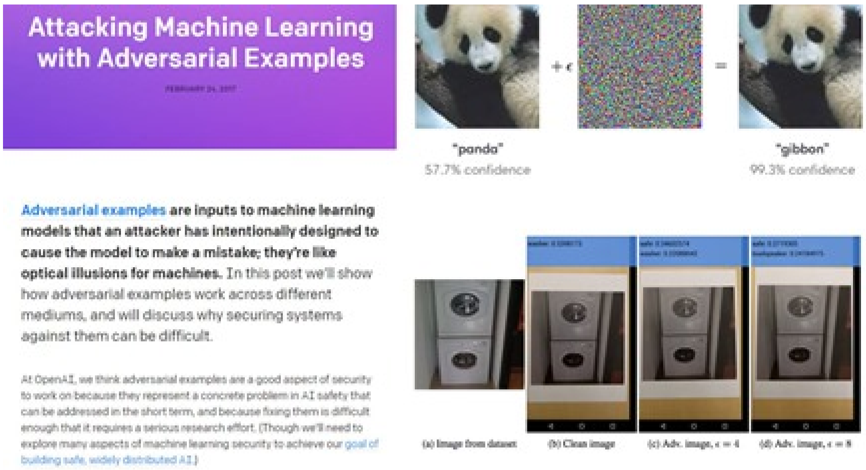

J'expliquais aussi ce que l'on appelle le "biais des données", et comment celui-ci peut amener à créer un système d'IA qui ne fonctionne pas bien dans la pratique lorsque ses données d'entrainement ne sont pas représentatives de son périmètre d'usage. L'exemple classique étant le système de reconnaissance faciale qui n'a été entrainé qu'avec des visages de blancs et qui ne reconnait donc pas les visages de couleur. Mais cet exemple, marquant, peut en illustrer d'autres.

Au-delà de considérations humaines et éthiques, la question qui se pose est purement statistique et probabiliste : un système d'IA à base de machine learning et de deep learning ne fonctionnera bien que si les données qui ont servi à l'entrainer sont représentatives des données qui seront ensuite exploitées en production. C'est exactement le même problème que dans un sondage politique : l'échantillon utilisé doit être représentatif de la population d'ensemble. Si vous faites un sondage uniquement à Paris, dans le XVIe arrondissement, dans le 93, en Mayenne ou à Strasbourg, vous n'aurez pas un échantillon représentatif de la diversité du pays.

De nombreux systèmes d'IA sont entrainés avec des sources de données internes et externes à l'entreprise. C'est un moyen de créer des modèles pertinents avec des données complémentaires, mais cela présente aussi une surface d'attaque plus grande par des cyberpirates. Il faut donc en tenir compte dans la conception des modèles.




Enfin, j'ai fait un rapide tour de quelques nouvelles menaces liées au machine learning et surtout au deep learning. Montrant par exemple comment on peut tromper des systèmes de vision artificielle en altérant très légèrement les images. Légèrement mais suffisamment pour modifier la reconnaissance d'images qui, dans la pratique, s'appuie sur des modèles probabilistes. Cf **Understanding the limites of deep learning** de Mariya Yao (mars 2017).




J'ai aussi illustré cela avec la méthode de création de masques permettant de tromper le login FaceID de l'iPhone X. Cette méthode digne de la série et des films Mission Impossible, consiste à créer un masque 3D imitant la forme d'un visage standard, sur lequel sont plaquées les photos imprimées en 3D des yeux, de la bouche et du nez de la personne dont on veut débloquer le téléphone. Pourquoi ces parties du visage ? Parce que ce sont celles qui sont reconnues par le système de login d'Apple, et par les systèmes d'identification faciale en général. C'est là que se situent les traits qui sont transformés en paramètres mathématiques uniques permettant de reconnaitre un individu. C'est une belle vulnérabilité, mais le hack n'est tout de même pas facile à réaliser. Seuls les pirates réellement déterminés l'exploiteront et sur des cibles dites "à forte valeur".


J'ai aussi présenté quelques exemples de "fake news" créés par des réseaux de neurones génératifs, comme le projet "**face2face reenactment**" qui permet d'animer un visage donné à partir de la captation de la vidéo d'une tierce personne, et celle consistant à faire parler Barack Obama avec un texte préparé. La technique utilise deux réseaux de neurones génératifs : l’un pour animer son visage et l’autre pour générer sa voix, originaire de la startup **Lyrebird**. Le tout est assez bluffant. Au passage, un petit truc : ces réseaux génératifs fonctionnent pour l’instant à basse résolution. Les images générées dépassent rarement 256 pixels de côté pour des raisons de puissance machine disponible. Mais avec les progrès du matériel, les faux vont probablement gagner rapidement en résolution et en réalisme.

Face2Face: Real-time Face Capture and Reenactment of RGB Videos

Juste après moi, j’ai assisté à une très bonne présentation pratique d’**Anaël Beaugnon** de l’**ANSSI**, l’agence du gouvernement qui gère la sécurité des systèmes d’information de l’Etat mais prodigue aussi des recommandations aux entreprises. Elle expliquait comment les solutions à base de machine learning étaient utilisées pour détecter des intrusions. Elle mettait bien en évidence les modèles statistiques sous-jacents et leurs limites. L’enjeu de ces systèmes est de minimiser leur détection de faux positifs ou, surtout, de faux négatifs, en raison d’effets de bord. Anaël Baugnon présentait l’outil SecuML développé à l’ANSSI qui sert à diagnostiquer un classifieur de cybersécurité. Il est en **open source sur Github.** J’ai au passage découvert de nombreux outils du monde de la cybersécurité en entreprise : la base de données de malwares **Contagio**, les outils de détection de fichiers PDF malveillants, le **projet ILAB** d’annotation de données par les experts pour la détection d’intrusion dont Anaël Baugnon est coauteure, ainsi que le projet **ALADIN**.

Sa présentation faisait bien écho à une tendance que l’on peut observer du côté des startups du secteur, qui se sont lancées dans l’utilisation du machine learning (avec ou sans réseaux de neurones) pour détecter des virus, phishings et autres intrusions. La tendance lourde du secteur consiste à créer des modèles statistiques qui détectent des “patterns” dans les logiciels suspects plutôt que de créer des bases de signature à la main. Ces systèmes exploitent essentiellement des techniques de machine learning.

Dans mon dernier ebook “**Les usages de l’intelligence artificielle**” publié en octobre 2017 (gratuit, et déjà téléchargé plus de 14 500 fois), je faisais ainsi un inventaire page 177 de quelques startups de ce secteur utilisant des modèles de machine learning : “Les tentatives de phishing sont détectées par **GreatHorn** (2015, $8,83m) ou avec **Lookout** (2007, $282m) qui sécurise les mobiles avec un modèle prédictif. Les malwares sont détectés avec du machine learning par **Cylance** (2012, $177M). L’israélien **DeepInstinct** (2014, $32M) 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. Dans le même genre, **Recorded Future** (2009, $33M) utilise le machine learning pour détecter les menaces de sécurité en temps réel. Des startups comme **Onfido** (2012, $30M) vérifient l’identité de clients de service en ligne. C’est de la détection de fraude basée sur du machine learning et du prédictif. L’israélien

**Fortscale** (2012, $32M) 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 autorisée !”.

Il faudrait aussi ajouter les outils qui permettent d’analyser les failles de sécurité de ses propres solutions logicielles. Les offres sont abondantes dans ce domaine. On peut notamment citer le framework open source **Frame-C** développé par le **CEA-LIST**, l’institut de recherche sur les systèmes numériques intelligents du CEA, qui s’appuie directement sur des méthodes formelles à base d’IA.

Dans la pratique, les modèles doivent cependant être entraînés par des bases de données associant de nombreux logiciels et leur niveau de menace. Vincent Letoux d’**Engie** insistait dans sa présentation sur le besoin de ne jamais évacuer l’Homme des processus de décision dans ces systèmes. C’est un point de vue de bon sens. De son côté, Dimitri Tromboff de **Thales** expliquait comment le machine learning était utilisé dans la pratique.

D’autres interventions portaient sur des sujets de cybersécurité éloignés de la thématique de l’IA, comme un post-mortem de la lutte contre les ransomwares. Il y avait notamment celle du dynamique **Eric Barbry**, avocat du cabinet d’Alain Bensoussan que l’on ne présente plus. Il expliquait en 45 minutes bien denses les enjeux de l’entrée en vigueur de la directive RGPD le 25 mai 2018, portant sur la protection des données personnelles et la vie privée. Les entreprises vont avoir du pain sur la planche ! Il existe d’ailleurs une véritable interaction entre la RGPD et l’IA, notamment autour des notions du droit à l’oubli et de celui de la portabilité des données. Lorsque vos données personnelles ont servi à entrainer un réseau de neurones, comment fait-on pour faire oublier votre existence à ce réseau de neurones ? La question est mathématique et logique autant que juridique !

Bref, comme dans tous les métiers du numérique, la cybersécurité n’échappe pas à la vague de l’IA. L’IA amène à revoir les raisonnements, elle change les méthodes, et permet de se remémorer que l’IA ne fonctionne pas toute seule. Elle est alimentée et corrigée en permanence par des données d’origine humaine.

Cet article a été publié le 30 novembre 2017 et édité en PDF le 15 mars 2024.