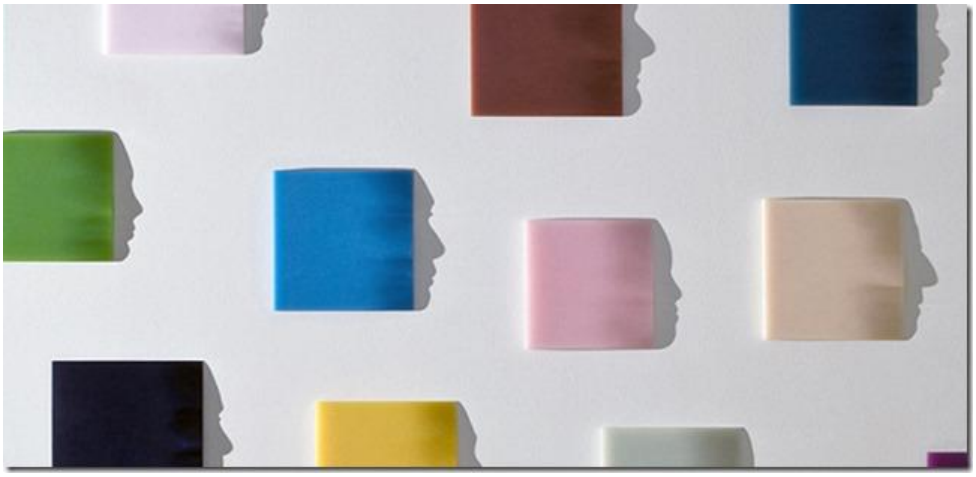

Kumi Yamashita - Portraits d'ombre

L'IA par ses ombres

Catégorie : **IA & Neurosciences**
Tags : **éthique, politique, cyberpolitique, technique, psychologie, société**
Personnages : **M.I. Jordan, T.M. Mitchell, Elon Musk, John Perry Barlow**

3 mars 2018

La lutte contre les usages malveillants de l'Intelligence Artificielle n'est-elle pas un prétexte de plus pour renforcer l'emprise des géants d'Internet ?

Préambule

Le mois dernier est parue une étude portant sur l'IA et ses usages potentiellement malveillants, étude qui a été abondamment relayée par la presse : « *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* ».

Nous nous appuyons sur ce travail pour proposer une série de commentaires constituée de deux parties.

La première partie porte sur quelques extraits choisis de l'étude et qui permettent de comprendre certains aspects importants de l'IA en 2018 : les prouesses accomplies par cette technologie en matière de « sens artificiels » (vision, sons...), la distance qu'elle finit par mettre entre nous et les conséquences de nos actions, sa capacité à individualiser en masse les services à moindre coût, la facilité avec laquelle nous pouvons la mettre en œuvre...

La seconde partie propose des commentaires plus « politiques » de l'étude. Car en abordant le sujet des usages malveillants de l'IA, il faut inévitablement prendre pied sur le terrain de la morale et de l'éthique, avec sa propre culture et sa propre subjectivité. Nous proposons donc d'examiner le point de vue des auteurs, d'où ils émettent leur avis et leurs recommandations. Ce terrain de l'éthique mérite à notre avis d'être occupé et

abordé selon d'autres angles et par d'autres acteurs de la vie publique, y compris nous-mêmes malgré, parfois, notre manque de « technicité ».

Technique

L'IA sort de l'ombre

Deux universitaires américains cités dans l'étude, M.I. Jordan et T.M. Mitchell, rappellent pourquoi les TIA (Techniques d'Intelligence Artificielle) sont sorties d'un long coma :

> Les principaux facteurs qui permettent d'expliquer ces progrès récents sont : la croissance exponentielle de la puissance de calcul, l'amélioration des algorithmes de « machine learning » (particulièrement dans le domaine des réseaux de neurones profonds), le développement de frameworks de logiciels standardisés qui permettent des itérations plus rapides et la reproduction des expériences, les ensembles de données [datasets] plus grands et plus accessibles et le renforcement des investissements commerciaux.

Les TIA ne sont pas nées d'hier et leur décongélation résulte quasi exclusivement de facteurs d'échelle (puissance des ordinateurs, données à profusion, investissements massifs…), d'organisation et de convergence (startups, frameworks, couplages TIA / machines…). Pas de nouvelle théorie ni d'Einstein de l'esprit artificiel mais le résultat est là : les TIA se répandent maintenant dans des infrastructures surpuissantes et miniaturisées en suivant nos données, nos images, nos voix... nos traces numériques.

Mais de quoi sont-elles aujourd'hui capables ?

Des sens artificiels in/sur-humains

Les avancées les plus spectaculaires concernent le traitement et la reconnaissance d'image (identifier quelqu'un / quelque chose sur une photo ou une vidéo) où les performances humaines ont été dépassées depuis 2016.

Techniquement, il n'est déjà plus possible d'échapper aux caméras de surveillance. Même en courant, notre visage peut être automatiquement reconnu, avec pour conséquence la réalisation possible d'un fantasme orwellien : la surveillance de masse[1]. Les TIA munissent le monde numérique de millions d'yeux de lynx (voire mieux puisqu'elles peuvent théoriquement prendre en charge tout le spectre électromagnétique et plus généralement tous les signaux physiques échappant à nos sens).

L'étape suivante consiste à *interpréter* une scène dans laquelle les personnages et les éléments de l'environnements ont été identifiés. « Comprendre » les scènes de la vie

1 Angélique Forget pour France Culture (petit podcast de 4 minutes) – 9 janvier 2018 – *En Chine, la cybersurveillance par la reconnaissance faciale*

réelle, ce qui se passe, est une compétence incontournable pour renforcer l'enserrement digital[2].

Les TIA permettent également de créer des images et des vidéos tout à fait bluffantes, avec côté lumière (en anglais)[3] :

[youtube https://www.youtube.com/watch?v=ttGUiwfTYvg?rel=0]

Et côté ombre des hackers déjà capables du pire[4].

Mais ce qui est possible avec les images l'est tout autant avec les sons ou le langage... Les TIA nous exposent donc à un déferlement de « fakes » humainement impossibles à déceler (sauf éventuellement à disposer nous-mêmes de sens augmentés par IA). Citons encore la reconnaissance de la parole (« OK Google »…), la « compréhension » du langage, les jeux (Échecs, Go, jeux vidéo...), la capacité des TIA à élaborer des comportements stratégiques, le pilotage de véhicules…

Il faut s'y résoudre : l'IA proposait déjà des « sens digitaux » au-delà de toute capacité humaine. Les TIA arraisonnent désormais le monde réel en dotant les infrastructures et les plateformes des acteurs du numérique de sens dignes du meilleur du monde animal.

L'IA, c'est facile !

On pourrait penser que tout ceci est très compliqué. Pas vraiment, et c'est bien le problème. Nous alertions récemment sur une faiblesse majeure des capacités françaises (Rapport #FranceIA : point et contrepoints) :

> […] nos forces et nos moyens résident principalement dans les mathématiques et l'algorithmique, dont les résultats sont… publics et open source, c'est-à-dire joyeusement partagés.

Et les auteurs de l'étude ne manquent pas de souligner que les technologies d'IA ne constituent une barrière d'entrée pour personne :

> En vérité, de nombreux algorithmes d'IA récents sont reproduits en quelques jours ou en quelques semaines. De plus, la culture de la recherche en IA est caractérisée par un haut niveau d'ouverture et de partage, avec beaucoup d'articles accompagnés de code source. […] on peut facilement trouver des algorithmes open

2 Elsa Trujillo pour Le Figaro – 25 octobre 2017 – *Google entraîne son intelligence artificielle à interpréter les gestes humains*

3 Au passage, voici bonne illustration de la différence entre Techniques d'IA (TIA), qui n'ont rien à voir avec l'intelligence, et l'IA, terme sujet à toutes les interprétations. Le bricolage d'images existe depuis longtemps (Photoshop...), mais les TIA permettent désormais, grâce à la puissance des ordinateurs, d'opérer ce bricolage sur une vidéo, en temps réel et de façon hyperréaliste. Indubitablement il s'agit d'un progrès technique, mais pas plus.

4 Elodie pour le Journal du Geek – 25 janvier 2018 – *Grâce à cette IA, tout le monde peut figurer dans un film porno*

source de détection de visages, de navigation et de planification, des frameworks multi-agents d'intelligence distribuée [5] pouvant être exploités à des fins malicieuses.

La complexité technologique ne met donc pas les TIA hors de portée des utilisateurs malveillants. De plus, en matière de cyberpiratage, nous savons bien que les chaînes de valeur sont parfaitement découpées et que chacun a son métier : un ransomware se vend et s'achète, comme demain (ou peut-être déjà aujourd'hui) des IA malicieuses, configurables et parfaitement documentées, seront disponibles sur les étals du darknet.

Individualisation de masse

Le « phishing » est une technique d'attaque bien connue visant à soutirer de l'information ou à initier une action malveillante en leurrant sa cible avec une façade digne de confiance : un faux mail, sms, tweet... et désormais, grâce aux TIA, une fausse vidéo, une fausse voix, etc.

Pour atteindre un niveau de confiance suffisant pour que la cible se dévoile, il faut collecter de l'information à son sujet (nom, amis, activité professionnelle, centres d'intérêt, sites visités, trainées GPS… - toutes sortes de traces numériques) et fabriquer un « masque crédible ». La difficulté pour le hacker réside dans l'équation coûts / bénéfices : il peut être long, coûteux et peu rentable de récolter de l'information et d'assembler un message crédible pour viser une seule personne.

Les TIA permettent désormais d'individualiser le phishing en masse et à moindre coût, depuis la collecte d'informations (senseurs digitaux et réels) jusqu'à la fabrication du message (texte, audio et/ou vidéo) dans la langue du destinataire.

Retenons ceci : le spear phishing (« intelligent », donc) est l'ombre portée de l'individualisation de masse par les TIA, qui permet de générer de la confiance et de l'empathie à moindre coût, un rêve marketing dans une société de l'individu.

Distance psychologique

Les activités assistées par l'IA se multiplient. Nos « appendices intelligents » (assistants personnels, robots…) se déploient et nous « augmentent » par leurs senseurs et actionneurs surhumains.

Dès lors, la « distance psychologique » entre humains s'accroit et peut s'étendre jusqu'au point où attaquer (au sens large) ne pose plus d'états d'âme, où l'agression n'est plus entravée par notre éducation, notre culture ou notre histoire. La victime est tout simplement virtuelle, voire totalement « prise en charge » par nos appendices immédiats ou distants (drone…).

[5] Il s'agit d'une approche basée sur une multitude d'agents simples dont le comportement global émergent présente des caractéristiques complexes, à l'image d'une colonie de termites ou de fourmis par exemple.

Nous serions donc, de ce fait, surexposés à des attaques a-morales, purement techniciennes. Peut-être, mais si c'est le cas, alors ce caractère induit par la distance s'applique aussi côté lumière : économie, société, politique...

Résonance technologique

Le phénomène de « résonance technologique » (Manifesto) se révèle ici comme une combinatoire de possibilités. Autrement dit, les formes d'attaques possibles explosent.

[youtube https://www.youtube.com/watch?v=RYzn_gmFs5w?rel=0]

Prenons un « corps » autonome : véhicule, drone (de livraison…), robot... Dotons-le d'outils de reconnaissance d'image, de son, d'interprétation de scène... Nous comprenons dès lors qu'il est possible de procéder à une attaque physique sans aucun contrôle à distance (grâce aux TIA, l'autonomie des artefacts s'étend progressivement en temps (plus longtemps) et en espace (plus loin). De plus, l'interopérabilité des « corps » et des algorithmes permet d'envisager la mise au point d'assemblages et de kits rapides, peu coûteux.

Citons à nouveau l'étude :

> En ce qui concerne les systèmes cyber-physiques, l'Internet des Objets (IoT – « Internet of Things ») est souvent annoncé comme une grande avancée en matière d'efficacité et d'utilité, mais il est aussi reconnu comme très fragile et il représente un autre vecteur d'attaque par lequel des systèmes d'IA contrôlant des systèmes-clés pourraient être sabotés, occasionnant plus de dégâts qu'il n'aurait été possible si ces systèmes avaient été contrôlés par des humains.

Ainsi, les menaces *physiques* suivent parallèlement les capacités physiques des TIA, assemblées en réseau ou couplées à des corps.

Politique

Qui parle ?

En évoquant ces quelques caractéristiques de possibles attaques, nous comprenons un peu mieux les capacités des TIA aujourd'hui. Que faire pour nous protéger de leurs usages malveillants ?

Les auteurs de l'étude font quelques propositions, mais tout d'abord qui sont-ils ? Pour la plupart, il s'agit de chercheurs provenant d'organisations anglo-américaines bien connues :

- Le « *Future of Humanity Institute* » et le « *Centre for the Study of Existential Risk* », deux institutions (déjà citées dans Un futur sans nous) militant activement pour participer au jeu politique ;
- Les universités d'Oxford et de Cambridge et le « *Center for a New American Security* » (CNAS) ;

- L' « *Electronic Frontier Foundation* », organisation active de défense des libertés sur internet, dont l'un des cofondateurs, John Perry Barlow, décédé le mois dernier, est connu pour avoir rédigé en 1996 la célèbre « *Déclaration d'indépendance du cyberespace* ».
- « *OpenAI* », une association à but non lucratif coprésidée par... Elon Musk.

Les observations et recommandations émanent donc d'un réseau de chercheurs et d'activistes plutôt enclins, comme nous allons le voir, à « prendre les choses à main ».

Sécurité politique

Voici le point de départ :

> L'IA favorise des changements dans la nature de la communication entre les individus, les entreprises et les états, de telle façon que tous sont mis en relation par des dispositifs automatisés qui produisent et présentent du contenu.

La pervasivité d'Internet permet le développement d'un nouvel environnement informationnel, dans lequel les (nombreux) effets ne sont plus attribuables à des causes claires. La raison classique devient inefficace, notre rapport à la « vérité » moins direct. C'est donc toute la production de contenu hors sol qui est désormais susceptible d'être attaquée, altérée, manipulée. Par qui ? Par quoi ?

> […] les technologies des réseaux sociaux sont à la disposition des autorités en place aussi bien que des protestataires : elles permettent au renseignement militaire de surveiller les sentiments et les attitudes […] certains prétendent que les réseaux sociaux ont contribué à polariser le discours politique en permettant aux utilisateurs, occidentaux en particulier, d'auto-sélectionner leurs propres chambres d'écho, mais d'autres mettent en doute cette hypothèse. Les algorithmes d'apprentissage automatique [machine learning] qui tournent sur ces plates-formes priorisent les contenus que les utilisateurs sont censés aimer.

Par conséquent, les TIA spécialement conçues pour adapter votre fil d'information, demain, pourquoi pas, de rédiger des tweets et des messages à votre place, etc. permettent d'envisager, moyennant un gros travail et une solide intelligence, des manipulations de masse (cela a-t-il déjà été le cas lors des dernières élections américaines ?).

> Les régimes autoritaires en particulier peuvent profiter d'un paysage d'information où la vérité objective (*objective truth*) est dévaluée et où la « vérité » est tout ce que les autorités déclarent comme telle.

Voici le terrain préparé pour insinuer un doute concernant les « autorités » et enfoncer le clou...

Réponse politique

... avec le marteau de la pure logique :

> […] les réponses appropriées [à toutes ces menaces] sont entravées par deux facteurs qui se renforcent l'un l'autre : d'un côté, le manque de compréhension technique de fond de la part du législateur, conduisant potentiellement à des réponses réglementaires, légales ou politiques mal conçues. De l'autre côté, la réticence des chercheurs à s'engager sur ces sujets, sans considération que l'association de leur domaine à des usages malveillants pourrait ternir la réputation de tout leur champ de recherche, conduisant ainsi à des dimensions de moyens ou à des réglementations prématurées.

Nous (re)touchons ici l'argument récurrent, que nous contestons à nouveau, pour que les scientifiques prennent seuls en charge l'éthique de leur domaine :

> […] il est essentiel que les chercheurs considèrent de leur responsabilité de faire tout ce qui est possible pour promouvoir les usages bénéfiques de la technologie et prévenir leurs usages néfastes.

Comment dire le contraire ? Il n'est pas étonnant d'observer le même début de phénomène en France (avec les neurosciences par exemple). Mais cette fois, l'argument consistant à dire « les politiques n'y comprennent rien donc les scientifiques doivent s'en mêler » est assené sur un endroit douloureux : celui de la sécurité. Imparable !

John Perry Barlow

Sans que ce point de vue extrême soit partagé par tous les auteurs du rapport, faut-il rappeler ce qu'écrivait John Barlow en 1996, et qui subsiste en arrière-plan :

> Gouvernements du Monde Industrialisé, vous, géants épuisés de chair et d'acier, je viens du Cyberespace, le nouveau foyer de l'Esprit. Au nom du futur, je vous demande, vous venant du passé, de nous laisser tranquilles. Vous n'êtes pas les bienvenus parmi nous. Là où nous sommes, vous n'avez aucune souveraineté.

Winners take even more…

Voici enfin un point important tiré de la conclusion de l'étude :

> En l'absence d'efforts significatifs, déterminer l'origine des attaques et sanctionner les agresseurs sera probablement difficile, ce qui pourrait conduire à une succession permanente d'attaques de faible et de moyenne intensité, érodant la confiance dans nos sociétés, entre les sociétés et leurs gouvernements et entre les gouvernements eux-mêmes.

Si la vérité est question d'intersubjectivité, alors semble plus « vrai » ce qui est « garanti pour vrai » par le plus grand nombre. Alors, en matière de confiance « *Winners take all* », comme on dit. Par conséquent :

> Les géants de la tech et des médias peuvent devenir des havres de paix technologiques pour les peuples car, par leur accès à des données pertinentes sur une échelle massive, par leur maîtrise des produits et des canaux de communication [et des infrastructures techniques sous-jacentes – Trevor Paglen le long des câbles sous-marins], ils sont dans la position hautement favorable de pouvoir offrir à leurs usagers la protection adéquate.

Par conséquent, les GAFAM deviennent encore plus nécessaires.

La question du rôle des gouvernements est évidemment centrale et des options politiques sont esquissées :

> Les états seront sous pression pour protéger leurs citoyens et préserver leur propre stabilité politique face à l'usage malveillant de l'IA. Ils pourront alors vouloir contrôler directement les infrastructures digitales [datacenters…] et de communication [câbles, satellites…], ou bien élaborer des collaborations fructueuses avec les entités privées qui contrôlent ces infrastructures, ou encore mettre en place une législation judicieuse et contraignante articulée avec des incitations financières bien conçues […]

Autrement dit, nous devons, nous experts, chercheurs, entrepreneurs, prendre les devants et agir auprès du pouvoir politique pour éviter cette « catastrophe » : que les états se mêlent de vouloir contrôler les infrastructures (dans une démocratie réellement transparente ce serait, du point de vue du citoyen, évidemment la meilleure solution…) !

Ce point étant assez délicat, la conclusion du rapport n'étonnera personne :

> Il subsiste beaucoup de désaccords entre les co-auteurs de ce rapport, sans parler des nombreuses communautés d'experts de par le monde.

Nous aurions vraiment aimé les connaître, ces désaccords !