Séance du 24 octobre 2022

L'intelligence artificielle va-t-elle contribuer à transformer *Homo Sapiens* en *Homo Numericus* ?

Michel CHEIN
Professeur émérite à l'Université de MONTPELLIER
Académie des Sciences et Lettres de MONTPELLIER

MOTS CLÉS
Intelligence artificielle, apprentissage automatique, systèmes d'Aide à la décision, neutralité et biais, risques, identité numérique, éthique

RÉSUMÉ
Intelligence Artificielle étant un terme polysémique, on utilise *IA* non pas comme un acronyme mais comme le nom d'un domaine de la science informatique.

On explique rapidement, à partir du problème de la reconnaissance de chiffres manuscrits, le principe de l'apprentissage automatique par réseaux de neurones, technique à la base des systèmes actuels d'IA.

L'IA, qui a développé des merveilles techniques, est plutôt stupide que géniale si on se réfère à l'intelligence humaine, on se demande alors pourquoi s'inquiéter.

Nous décrivons des erreurs dans des systèmes d'Aide à la Décision concernant des personnes (médecine, police, justice, recrutement, demande de prêt, …), erreurs dues aux biais inévitables dans les données car une personne est réduite à un petit ensemble de nombres.

L'un des risques majeurs est qu'un utilisateur d'un tel système est amené, pour l'utiliser au mieux, à se conformer au système, c'est-à-dire à identifier une personne insidieusement à un avatar numérique.

Quelques considérations éthiques et quelques pistes pour éviter la catastrophe que serait la transformation Homo Sapiens en Homo Numericus et l'ordinateur, ou l'IA, en un nouveau veau d'or sont données en guise de conclusion.

Nota : Cette conférence a auparavant été donnée par l'auteur à la Conférence Nationale des Académies qui s'est tenue à Caen du 5 au 8 octobre 2022. Le texte en a été publié dans les « Mémoires de l'Académie des Sciences, Arts et Belles-Lettres de Caen, Tome LVII-2022, pp 77-108 ». Nous remercions vivement l'Académie des Sciences, Arts et Belles-Lettres de Caen pour nous avoir autorisés à reproduire ce texte dans le présent bulletin.

Le titre de cette conférence : *Intelligence artificielle, neurosciences, biotechnologies : vers quel Humanisme ?* pose lui-même de nombreuses questions. Des technologies – Ô combien importantes puisqu'elles concernent la vie, mais quid des sciences de la vie ? –, des sciences – Ô combien importantes puisqu'elles concernent,

entre autres, le cerveau, mais quid des technologies neuronales ? –, un domaine – l'Intelligence artificielle (IA dans le texte) objet de nombreuses craintes et de tout autant d'espoirs, dont on parle tous les jours avec des acceptions différentes, difficile à identifier science, technique, sorcellerie ? – et l'informatique et le numérique ? Et pourquoi ces domaines conduiraient-ils à un seul Humanisme, ne faudrait aussi s'interroger sur les humanismes possibles et quels sont les choix, en particulier politiques, permettant, si ce n'est d'aller vers un « avenir radieux » du moins d'éviter le pire… Dans l'introduction j'explique les limitations drastiques – ne considérer que le domaine de l'aide à la décision qui fait l'objet des deux parties centrales de ce texte – qui m'ont semblé pertinentes pour aborder ce sujet immense (toute la société est concernée, de l'individu à l'avenir de la planète), passionnant (quelles limites, jusqu'où pourrait-on aller ?), inquiétant (du contrôle des applications à celui de la population), complexe (qui a une vision du Web 3 ?) et urgent (la technologie court plus vite que les débats) de cette conférence. En conclusion, je propose quelques recommandations pour éviter le pire.

Introduction

Inutile d'insister sur les aspects positifs de nombreuses applications de l'IA, en santé – par exemple dans l'analyse d'images, l'analyse des EEG, le monitoring, la robotique médicale –, en recherche d'informations et recommandation, pour traquer le harcèlement en ligne, calculer des itinéraires en utilisant l'état du trafic, dans la conduite d'avion, les anti-spams, ces applications sont très présentes dans les médias.

Inutile également de dénoncer certaines pratiques bien connues comme l'hameçonnage personnalisé (c'est-à-dire la détermination de cibles intéressantes et la génération de messages crédibles en fonction des profils de ces cibles) ; les atteintes aux droits de l'homme, en particulier à la vie privée ; tout ce qui conduit à une véritable perversion du débat démocratique, les manipulations politiques par des campagnes de propagande automatisées et personnalisées, en utilisant des données personnelles vendues ou volées (il est bien connu maintenant que dans la campagne de Donald Trump, la société Cambridge Analytica détermina des États pouvant basculer, puis dans ceux-ci des électeurs incertains, et enfin envoya des messages ciblés, le *Brexit* aurait été entaché lui aussi par ce type de manipulation et les élections politiques doivent maintenant être protégées contre ces attaques) ; la fabrication de *fake new*s en utilisant des fausses photos ou la manipulation de vidéos avec des systèmes permettant la synthèse de la parole. Tous ces dangers sont renforcés parce que les techniques de l'IA sont un monopole des géants de l'informatique qu'ils soient américains (GAFAM) ou chinois (Alibaba, Baidu, Tencent, Xiaomi etc.) et lorsqu'elles sont aux mains d'un régime autoritaire ou totalitaire comme en Chine, elles pourraient conduire rapidement à un contrôle total de la population et à la construction d'un « homme nouveau » (ce que décrit avec angoisse Kai Strittmatter [1]).

Cependant, on peut se demander si, dans un pays démocratique, il est nécessaire de s'inquiéter, c'est-à-dire se demander si les techniques d'IA sont porteuses, en elles-mêmes, d'un tel danger, si l'utilisation massive de systèmes intégrant de l'IA pourrait contribuer à transformer l'Homme en un *Homo Numericus* premier pas vers un *brave sujet décervelé* ? Les techniques d'IA sont-elles si puissantes que cela ?

[1] Dictature 2.0 *Quand la Chine surveille son peuple (et demain le monde)*. Tallandier, 2020.

Schémas de Winograd

Auguste, le clown, essaie depuis plusieurs minutes de mettre une immense statue dans une minuscule valise. Monsieur Loyal, qui observe ses nombreuses tentatives suscitant les rires du public, lui dit au bout d'un moment : « Mais arrêtez donc, vous voyez bien que la statue n'entre pas dans la valise, elle est trop grande ! » Auguste s'arrête, regarde attentivement, l'air étonné, la statue et la valise, sort un long mètre jaune pliant en bois de sa chaussure, mesure la statue et la valise, recommence, lève les yeux au ciel et finalement, sûr de lui, lui répond « mais non, elle est trop petite ! »

Depuis Turing et son jeu de l'imitation[2], l'écriture d'un programme de *questions/réponses* en langue naturelle est un problème scientifique qui a suscité de nombreux travaux mais qui n'a toujours pas de solution générale satisfaisante. On s'est aperçu que *le test de Turing* (le jeu de l'imitation dans sa version simple) conduisait à écrire des programmes consistant à tromper une personne plus qu'à apporter des réponses directes et claires aux questions posées, ceci en utilisant jeux de mots, plaisanteries, citations, débordements émotifs, apartés, etc. Pour expliquer des réponses étranges faites par un programme celui-ci se faisait passer pour un jeune Ukrainien de 13 ans maîtrisant mal l'anglais ! Très critiquée la *Loebner Competition* entre programmes prétendant passer le test de Turing s'arrêta en 2016. Winograd suggéra un cadre limité et précis pour ce problème, que Levesque proposa d'appeler schéma de Winograd.

Par exemple, le schéma associé au dialogue des clowns est le suivant :

Énoncé : *La statue n'entre pas dans la valise car elle est trop [grande/petite]*.
Question : *qu'est-ce qui est trop [grande/petite] ?*
Réponse : [*la statue/la valise]*.

Plus généralement, un tel schéma consiste en un énoncé avec une alternative, une expression spéciale entre crochets (ici grande ou petite), et une question concernant cette expression spéciale avec les contraintes suivantes :

- la question consiste à demander quel est le référent du pronom de l'énoncé (ici elle),
- deux réponses sont possibles suivant l'expression spéciale tirée au hasard (grande ou petite),
- répondre doit être très facile, quasi instantanée, pour des humains,
- la réponse doit faire appel à des connaissances sur le monde,
- on doit utiliser des raisonnements de bon sens,
- on ne doit pas pouvoir répondre en utilisant des méthodes statistiques ou des moteurs de recherche (on dit qu'elle doit être *google-proof*),
- l'évaluation est faite automatiquement en ne comptant pas de la même façon une bonne et une mauvaise réponse, les mauvaises réponses sont pénalisées[3].

Voici un autre exemple :

Énoncé : *Paul a essayé de joindre Georges sur son téléphone, mais* ***il*** *[n'a pas réussi / n'a pas répondu]*.
Question : *Qui [n'a pas réussi / n'a pas répondu]* ?
Réponses : [*Paul / Georges]*

Une compétition existe sur ce sujet depuis 2011 (à laquelle ni IBM ni Google, entre autres, ne participent) dont les résultats montrent bien les difficultés posées par un

[2] Par exemple voir Test de Turing dans Wikipedia.

[3] Pour N questions sur un domaine la note est : max (0, N – k. Nombre Réponses Fausses/ N), où $k \geq 2$ est un coefficient pénalisant les mauvaises réponses (par exemple avec k=2 si une réponse sur 2 est fausse la note est 0).

raisonnement de sens commun simple pour un humain (cf.[4] pour un ensemble de schémas de Winograd).

Les schémas de Winograd ne sont qu'un exemple de tâches intellectuelles simples pour un humain que la machine ne sait pas bien résoudre aujourd'hui – il en existe beaucoup d'autres, « un crocodile peut-il courir un 110 mètres haies ? » est un exemple célèbre de question à laquelle un programme ne sait pas répondre … tant qu'on ne lui a pas donné la réponse.

Le désastre des deux tentatives du robot de conversation (*chatbot*) *Tay* de Microsoft est un exemple bien connu dans le domaine du traitement automatique des langues de limite de systèmes utilisant des techniques d'apprentissage. Tay devait passer pour une adolescente, initialement décrite, en particulier, à partir de données publiques, puis évoluer en discutant sur Twitter. En quelques heures, les tweets de Tay sont devenus racistes, misogynes et antisémites ! Si le système de Microsoft avait sans doute bien choisi les données initiales il n'avait pas prévu les tweets des utilisateurs.

Même si des systèmes de traduction sont spectaculaires pour certains types de textes, l'IA a des progrès à faire dans le domaine du traitement automatique des langues ! Mais au fait, c'est quoi l'IA ?

Polysémie galopante

Le terme IA est surchargé de significations, ce qui fait que l'IA peut parfois apparaitre comme une espèce d'auberge espagnole où chacun apporte ce qu'il souhaite. En se limitant à ses usages scientifiques, alors qu'il y a des usages commerciaux, politiques, voire idéologiques, on peut constater quatre phénomènes qui expliquent cette polysémie galopante.

Un phénomène de restriction : l'IA c'est l'apprentissage numérique (notons que c'est l'usage dominant aujourd'hui dans la presse, l'économie et la politique).

À ce phénomène de réduction est associé un phénomène concomitant d'expansion, l'IA c'est toute l'informatique plus la robotique, plus l'aide à la décision, plus etc.

Il existe également un processus d'évaporation et ce depuis le début, des domaines s'autonomisent ou s'immergent dans l'informatique. Un exemple ancien est celui de la linguistique computationnelle, un exemple actuel est celui des sciences des données. Ce qui était jusqu'à ces dernières années une partie de l'IA devient une science autonome.

On assiste aussi à un phénomène de neutralisation de certains domaines de l'IA par l'informatique, qui en les absorbant, fait perdre à ces domaines leur étiquette IA. En effet, les chercheurs en IA sont des explorateurs de l'informatique. Ce sont des chercheurs en IA qui sont à l'origine de nombreuses notions devenues banales en informatique : le temps partagé, les interfaces homme-machine, les langages fonctionnels, les langages objets, le *backtracking*, l'alpha-beta *pruning*, le *garbage collector*, les systèmes multi-agents, la programmation logique, la programmation par contraintes, etc. Ce que Nick Bostrom (directeur du *Future Humanity Institute*) décrivait en 2006 de la manière suivante : « Beaucoup d'IA de pointe a filtré dans des applications générales, sans y être officiellement rattachée car dès que quelque chose devient suffisamment utile et commun, on lui retire l'étiquette d'IA. Aujourd'hui ce serait plutôt l'inverse ! Quand l'IA a le vent en poupe, comme en ce moment, tout le monde fait de l'IA parce que c'est excitant, à la mode, et c'est aussi pour bénéficier des financements des plans nationaux ou pour conquérir des marchés. »

Depuis environ 70 ans l'IA, en tant que discipline scientifique avec ses nombreux journaux scientifiques et congrès internationaux, avec ses sociétés savantes (tous les pays

[4] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.html

économiquement développés ont une telle société savante), s'est développée suivant les grands axes : Apprendre, Acquérir et Représenter des connaissances, Raisonner, avec des ruptures, des échecs et des succès.

L'IA faible, qui consiste à construire des programmes simulant des tâches intellectuelles spécifiques – un tel programme est souvent appelée *une* IA– a obtenu de nombreux succès.

L'IA forte, qui a pour objectif de construire un programme simulant toutes les capacités cognitives humaines, voire plus, n'est toujours qu'un projet, ce n'est qu'une idée, une conjecture qui fait travailler certains chercheurs et suscite de nombreux débats et controverses, ce qui est sûr c'est que ce n'est pas pour demain. Il n'y a pas que la beauté et les arts, les sentiments et la conscience qui soient hors du champ actuel de l'IA, personne ne compte sur une machine pour inventer une nouvelle théorie dans n'importe quel domaine scientifique. Et rappelons que la complexité des machines est limitée et que l'être humain est d'une complexité biologique (matérielle) beaucoup plus grande que celle des ordinateurs physiques (même quantiques… lorsqu'on les maîtrisera).

Existe-t-il des intelligences artificielles intelligentes ?

« *Il* ne comprend même pas ce qu'il fait », « *il* ne comprend même pas pourquoi il le fait », « *il* ne sait même pas que ça ne s'applique pas dans ce cas-là », autant d'expressions qui conduiront à penser qu'*il* n'est pas très intelligent. Difficile de dire qu'on est intelligent si on ne comprend pas ce qu'on fait, si on ne peut pas dire pourquoi on fait telle chose, pourquoi on utilise tel outil, quelles sont ses conditions d'usage. Fabien Gandon dans *Les IA comprennent-elles ce qu'elles font ?* explique que comprendre « c'est saisir le sens, les finalités, les causes et conséquences, les principes. Comprendre quelque chose, c'est recevoir ou élaborer une représentation de cette chose, c'est s'approprier une conceptualisation reçue ou construite, qui permettra notamment de produire un comportement intelligent. »

Dans le cas du *chatbot* Tay, ce que Microsoft n'avait pas envisagé, alors que des spécialistes des réseaux sociaux auraient pu le prédire, était qu'un groupe d'utilisateurs de Twitter commencerait immédiatement à tweeter des commentaires racistes et misogynes. Tay a rapidement appris et a incorporé ces commentaires racistes dans ses propres tweets parce qu'il ne comprenait pas le sens des mots.

Comprendre c'est être capable de répondre à une succession de *pourquoi ?* par des réponses de plus en plus précises correspondant à des niveaux de compréhension de plus en plus détaillés au fur et à mesure que des *pourquoi ?* s'enchaînent. Par exemple, différents niveaux de compréhension, d'explication, de la succession du jour et de la nuit, pourraient être, dans un premier temps : le soleil est une source fixe de lumière et la terre tourne sur elle-même, puis ajouter que la terre tourne autour du soleil, puis que l'axe de rotation de la Terre fait avec la perpendiculaire au plan de l'écliptique un angle constant de 23° 26′, puis introduire les variations de luminosité, de température et de composition atmosphérique de la Terre, etc.

Dans le titre de ce paragraphe « intelligence artificielle » est une locution désignant un programme, ou un ensemble de programmes utilisant certaines techniques et certaines données, alors que l'attribut « intelligent » est pris dans son sens humain habituel, si bien qu'on peut dire, aujourd'hui, qu'une intelligence artificielle n'est pas très intelligente voire que toutes les intelligences artificielles sont stupides puisqu'elles ne comprennent pas ce qu'elles font ni pourquoi elles le font !

C'est une personne qui fixe les objectifs du programme (jouer à un seul jeu ou à plusieurs, jouer le mieux possible ou avoir plusieurs niveaux de jeu, battre un champion du monde, avoir une interface graphique en 2 ou 3 dimensions, etc.) qui décide de sa

conception et de sa construction (quelle combinaison de différentes techniques : apprentissage automatique, recherche arborescente, fonction d'évaluation, Monte-Carlo, règles logiques, accès à des bases de parties, etc.), de ses conditions d'usage (essayer d'utiliser Deep Blue pour jouer au Go !), de son évaluation (peut-on imaginer Deep Blue sans Kasparov ou Alpha-Go sans Lee Sedol ?).

Et lorsque l'on parle d'un artefact « intelligent » c'est dans le sens où sa conception et sa réalisation ont nécessité beaucoup d'intelligence humaine ou qu'il a un comportement efficace pour une tâche limitée, à ce dernier sens de nombreuses IA sont souvent beaucoup moins « intelligentes » qu'un mécanisme d'horlogerie, et s'il existe, certes, des IA qui sont des programmes très complexes, c'est l'intelligence des hommes qui les ont réalisées.

Comprendre le sens d'un symbole, d'un mot en particulier, nécessite que celui-ci soit relié dans le cerveau à des zones cérébrales le concernant (cf. pour une présentation amusante le blog binaire[5]). Nils Nillson, chercheur éminent en Intelligence Artificielle récemment décédé, a écrit un article de titre « *Human-Level Artificial Intelligence ? Be Serious !* »[6] Gérard Berry, lui aussi chercheur éminent en informatique, disait en 2016 : « Je n'ai jamais été déçu par l'intelligence artificielle parce que je n'ai pas cru une seule seconde en l'intelligence artificielle. Jamais. », et, un peu plus tard, il précisait son point de vue en écrivant que l'IA est la « quête d'une intelligence non naturelle, implémentée sur un ordinateur par exemple. Certains disent que le sujet n'existera que quand l'ordinateur sera capable de pouffer de rire alors qu'on lui raconte une histoire drôle qu'il ne connaît pas. Mais ce sont des mauvaises langues. J'ai appris que Google et Facebook s'étaient cotisés pour construire le premier ordinateur réellement intelligent. Je connais la première question qu'ils vont lui poser : « Dieu existe-t-il ? » Et je devine la réponse : « Maintenant, oui. » Mais trêve de balivernes : pour les gens sérieux du domaine, c'est aussi une grande quête sur les possibilités réelles des machines comparées à nos propres limites, qui peut aussi nous éclairer sur ce que nous appelons sans trop de modestie l'« intelligence humaine »[7].

Mais alors… pourquoi s'inquiéter ?

Reconnaissant que les IAs sont plutôt stupides que géniales, même si elles peuvent être des merveilles techniques, pourquoi s'inquiéter ?

Il suffit de quelques titres de livres récents écrits par des Français (la littérature anglaise est beaucoup plus importante) – *L'Intelligence artificielle n'existe pas*. Luc Julia (informaticien, co-inventeur de Siri), First éd. 2019 ; *L'Intelligence artificielle ou l'enjeu du siècle. Anatomie d'un antihumanisme radical*. Éric Sadin (philosophe), L'Échappée, 2018 ; *La fabrique du crétin digital*. Michel Desmurget (neuroscientifique), Le seuil, 2019 ; *L'homme nu. La dictature invisible du numérique*. Marc Dugain (romancier, journaliste), Christophe Labbé (journaliste), Plon, 2016 – pour se dire que malgré les nombreux succès de l'IA, dont je ne parlerai pas car ils sont quotidiennement exposés dans les médias, les choses ne sont pas si simples.

Étant informaticien, spécialiste d'IA mais absolument pas philosophe, je limiterai drastiquement mon propos pour essayer de ne pas dire trop de bêtises. J'aborderai les liens entre IA et Humanisme en m'intéressant principalement à la question suivante : à quels effets anthropologiques pourraient mener, ou mènent déjà parfois, l'utilisation de

5 https://www.lemonde.fr/blog/binaire/2021/12/14/que-se-passe-t-il-dans-les-cerveaux-des-cons/

6 Ai Magazine, 25th Anniversary Issue

7 Gérard BERRY, *L'Hyperpuissance de l'informatique, Paris,* Odile Jacob, 2017.

certains logiciels d'aide à la décision utilisant des techniques d'IA et dont les objectifs concernent des humains – ou plus généralement le vivant – si leurs usages n'étaient pas rigoureusement contrôlés ?

Donc il ne sera pas tellement question de science, un peu de technologie mais surtout d'applications dans un domaine particulier. Cette sélection, en partie arbitraire comme toute classification, ferait sans doute réagir Philippe Descola, puisque d'une certaine manière elle calque la séparation entre les choses et l'humain, comme si les conditions de création, de distinction, d'usage des choses n'avaient rien à voir avec l'humain, d'autant plus que parmi ces choses j'inclus des artefacts ! Cependant, je pense que cette séparation permet de mettre en évidence, dans les applications de l'IA, ce qui tend à déshumaniser l'humain, à le transformer en un homme informatique ou numérique, oxymore s'il en est, du moins de mon point de vue.

Aide à la décision

La définition classique de l'aide à la décision, telle que celle proposée par Bernard Roy : « L'aide à la décision est l'activité de celui qui, en prenant appui sur des modèles, aide à obtenir des éléments de réponse aux questions que se pose un intervenant dans un processus de décision, éléments concourant à éclairer la décision et à recommander un comportement de nature à accroître la cohérence entre l'évolution du processus et les objectifs de cet intervenant », peut être généralisée à l'acception actuelle en considérant non plus seulement des humains mais aussi en utilisant des agents artificiels, des IAs (si on généralise aussi l'intervenant à des agents artificiels on obtient des systèmes automatiques ou robotiques dont je ne parlerai pas ici).

Par exemple, IBM décrit ainsi son système *Watson for Cybersecurity* utilisé pour la détection et l'investigation d'attaques : « la détection d'attaques se fait en corrélant les incidents, et l'analyse de sécurité peut solliciter l'aide de Watson pour comprendre l'origine du problème et les signaux remontés » et, complète Hugo Madeux, « la logique s'inscrit toujours dans le cheminement, l'enrichissement, l'analyse, le raisonnement ». Ainsi, la décision finale est prise par un humain et l'IA ne fournit que des outils d'aide permettant d'utiliser au mieux l'intelligence humaine. Mais les outils informatiques transforment, comme n'importe quel outil, ceux qui les utilisent. Les outils de l'IA étant des outils « intellectuels » ils transforment la manière de penser de ceux qui les utilisent et peuvent même parfois prendre leur place...

Donnons quelques exemples pour préciser les systèmes qui nous intéressent plus particulièrement ici.

Themis est un système comprenant un micro et un haut-parleur commandés par un logiciel de reconnaissance du langage qui, expliquerait un enfant, « crie » lorsqu'on prononce un « gros mot ». Naturellement, même si un tel système risque d'avoir des difficultés à reconnaître certaines figures de style il pourrait être utilisé pour apprendre à parler « correctement » une langue, correctement au sens du concepteur du logiciel !

À l'opposé, il existe des systèmes de sécurité qui « crient » lorsqu'ils anticipent une menace, par exemple lorsque la cuve d'un réacteur nucléaire risque de se fendiller.

Nous ne nous intéressons qu'à des systèmes dont les décisions concernent des humains et pas des objets. Ces systèmes peuvent être plus ou moins : incitatifs (e.g. systèmes de recommandation dans le marketing), prescriptifs (e.g. utilisés par exemple dans les banques ou les assurances) ou coercitifs (e.g. refus d'embauche, surveillance par la police ou condamnation par la justice)[8].

[8] Distinction proposée par Éric Sadin dans l'opus cité

La plupart des différents systèmes prédictifs dont on parle aujourd'hui utilisent les mêmes techniques d'apprentissage, particulièrement des réseaux de neurones profonds (*deep-learning*), sur des données massives (*big data*) annotées. Avant d'aller plus avant il est nécessaire d'expliquer très rapidement ce qu'est le *deep-learning*.

Apprentissage automatique par réseaux de neurones

Nous présentons le principe de ces méthodes (supervisées) à partir du problème de la reconnaissance de chiffres manuscrits. Ce problème, facile pour un humain est resté longtemps difficile pour un programme (les premiers *captcha* étaient composés de lettres ou des chiffres). Il suffit d'essayer d'écrire un programme reconnaissant des chiffres manuscrits pour s'en convaincre. Par exemple, pour décrire un « un » manuscrit on pourrait dire qu'il est *souvent* composé de deux segments, le plus long, *proche* de la verticale, a son extrémité supérieure *voisine* de celle du segment le plus court. Ces deux segments, dans un rapport *approximatif* de 1 à 3, font un angle *d'environ* 45 degrés ; ils peuvent être complétés par une base horizontale, de longueur *proche* de celle du segment le plus court, accueillant en un point *proche* de son milieu l'extrémité inférieure du segment le plus long. Certains « un » manuscrits sont aussi composés d'un unique segment, etc. Bref, les nombreux cas, les exceptions, les approximations font que ce qui est simple pour un humain est difficile à expliciter et à formaliser donc difficile pour un programme, sauf s'il est capable d'apprendre.

L'apprentissage supervisé consiste à considérer de nombreux exemples de chiffres manuscrits, et à améliorer les paramètres du programme jusqu'à ce qu'il les reconnaisse suffisamment souvent. Les récents succès pour résoudre ce type de problèmes ont été obtenus avec des réseaux de neurones artificiels. Un neurone artificiel simple fonctionne comme un mécanisme d'admission à un examen. On a une note par matière, chaque matière est affectée d'un coefficient, et une personne est reçue à l'examen si la somme des notes pondérées dépasse un seuil sinon elle est collée.

Un réseau de neurones comme celui de la Figure ci-dessous est composé de tels neurones, les sorties d'un neurone pouvant être reliées aux entrées d'autres neurones. Dans l'exemple de la reconnaissance des chiffres, on donne en entrées les valeurs des pixels d'une image d'un chiffre et on modifie les poids et les biais jusqu'à ce que pour un chiffre, par exemple « un », à force de fournir de nouveaux exemples de « un », le réseau ait une sortie toujours supérieure à 0.5 sur le neurone de sortie correspondant au chiffre « un », et pour les neuf autres neurones de sortie toujours une sortie inférieure à 0.5.

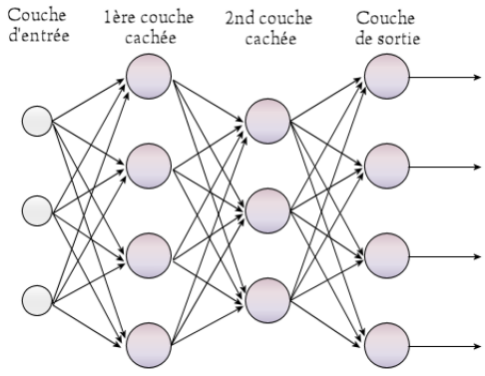

Figure : Un réseau de neurones

La phase d'apprentissage est terminée lorsque le réseau donne les bons résultats sur les exemples. Le réseau est alors évalué sur un autre échantillon de données et si les résultats sont satisfaisants (s'il donne suffisamment souvent la bonne sortie) on considère que le réseau a appris, on peut alors l'utiliser pour reconnaître des données quelconques de chiffres.

Les banques, pour reconnaître les chèques, et les postes, pour reconnaître les adresses, ont été les premiers grands utilisateurs commerciaux de ces techniques qui ont depuis été utilisées avec succès dans de nombreux domaines (reconnaissance d'images, de sons, de mots, dans des jeux, traduction, …). Trois raisons expliquent ces nouveaux succès : premièrement, l'informatisation de la société (internet, les réseaux sociaux, les *data centers*) permet d'avoir de très nombreux exemples ; deuxièmement, les machines de plus en plus rapides permettent de traiter ces grandes masses de données ; troisièmement, des algorithmes efficaces ont été inventés.

On vient d'expliquer très succinctement *comment ça marche* dans le cas d'un système apprenant à reconnaître un chiffre manuscrit, ou un animal, ou un bateau de pêche, ou une tumeur cancéreuse ou un blindé etc. On peut remarquer que les humains n'apprennent pas comme ça, un enfant n'a pas besoin de millions d'exemples d'images de chats pour savoir reconnaître un chat !

Reprenons les étapes d'une application basée sur une méthode d'apprentissage :

1. définir l'objectif de l'application (reconnaître des nombres ou des champignons, faire acheter des livres, jouer – univers fermé – ou conduire une voiture – univers ouvert, non seulement l'environnement mais aussi les conducteurs, déterminer un risque d'attentat, etc.) ;
2. choisir une méthode (e.g. apprentissage supervisé par un réseau de neurones) ;
3. acquérir des données d'apprentissage et les analyser : les données sont-elles pertinentes, c'est-à-dire représentatives du but (reconnaissance de chèques ou risque d'attentat) ? Si le modèle d'apprentissage choisi a beaucoup de paramètres il faut beaucoup de données.
4. évaluer sur d'autres jeux de données que ceux utilisés dans la phase d'apprentissage et qui doivent être représentatifs des cas réels qu'il faudra traiter.

Qui décide ?

Nous évoquons ici un problème fondamental : le glissement d'un système d'aide à un décideur humain vers un système automatique. Un système d'*aide* à la décision sous-entend que c'est un intervenant humain qui prend la décision. Le décideur devrait donc être capable d'expliquer comment il a pris sa décision à partir des propositions faites par le système donc comprendre, et expliquer, comment le système fonctionne, pourquoi il a fait de telles propositions et pourquoi lui, le décideur, a fait son choix dans cet ensemble. Tout le monde a eu l'occasion d'obtenir comme réponse à une question concernant une décision « je n'y peux rien c'est l'informatique ! » demain nous entendrons « je n'y peux rien c'est l'IA ! ». Tous les systèmes d'aide à la décision sont susceptibles de ce glissement. C'est un problème crucial : de tels systèmes d'aide à la décision doivent rester « d'aide » à des décideurs humains et ne doivent pas prendre le contrôle en particulier quand leurs décisions concernent des humains qui ne sont pas réductibles à un modèle informatique.

Un exemple qui illustre dramatiquement cela est celui du passage de drones télécommandés à des armes autonomes puisqu'il s'agit dans ce cas de décider de qui vit

et qui meurt (pour ce sujet largement commenté on peut regarder la vidéo fictive mais réaliste *Slaughterbots – if human : kill()*[9]) !

Certaines méthodes ne donnent comme explication que la méthode elle-même (dans l'exemple précédent de reconnaissance de chiffres, pourquoi le programme a-t-il reconnu un 7 et pas un 1 ? Parce que j'ai fait de nombreux calculs aboutissant à cette conclusion, pourrait répondre le programme. Un système à base de règles pourrait expliquer qu'il avait conclu par un 7 et pas par un 1 parce qu'il avait détecté une ligne brisée constituée de deux traits, l'un oblique 2 fois plus long que l'autre horizontal, la jonction se faisant entre l'extrémité nord du premier et est du second etc. Un directeur des ressources humaines utilisant un système d'aide au recrutement annonce à un candidat « nous sommes désolés : vous ne remplissez pas les conditions pour occuper ce poste. » Pourquoi, demande le candidat ? Le DRH devrait répondre : « Parce que les personnes que nous avons refusées (en fait, celles refusées dans les bases données utilisées pour l'apprentissage) avaient des paramètres semblables aux vôtres » ou bien « parce que les personnes que nous avons recrutées (ou recrutées ailleurs pour un tel poste) avaient des paramètres incompatibles avec les vôtres. » Dans les programmes basés sur de l'apprentissage automatique statistique, tout n'est affaire que de corrélations.

Neutralité et biais

Les moyens pour développer des applications étant nécessairement limités, et ceci quel que soit le domaine concerné (médical, politique, militaire, etc.) le but choisi pour le développement d'une application n'est pas neutre (dans toutes ces alternatives fictives, le *et* au lieu du *ou* exclusif serait souhaitable… maladie d'Alzheimer ou autisme ? manipulation de vote ou critiques au gouvernement ? menaces d'une grande puissance totalitaire ou terrorisme ?). Il n'y a pas non plus de neutralité dans le choix de la méthode, des données et de l'évaluation des résultats.

Parmi ces problèmes nous considérerons celui concernant les données. Avoir des données pertinentes, c'est-à-dire représentative du domaine concerné, est un problème fondamental de statistique c'est donc un problème fondamental des méthodes du type *deep learning*, et si les données concernent des humains leur analyse et leur évaluation sont du ressort des sciences humaines et sociales devant prendre en compte, entre autres, la manière dont elles ont été acquises (par achats, par butinage sur les réseaux, par *crowdsourcing*, extraites de bases de données expertes, etc.). Nous rappelons ci-dessous quelques exemples de biais dans les données utilisées par certains systèmes.

Médecine

Il ne s'agit pas ici de nier l'intérêt des systèmes l'analyse d'images (reconnaissance de certaines tumeurs par exemple), l'analyse des EEG, le monitoring, les systèmes prédictifs d'aide au diagnostic, comme récemment celui concernant l'insuffisance cardiaque[10] (de nature similaire à un système d'alarme sur une centrale nucléaire), mais de décrire les biais d'un système utilisé aux États-Unis identifiant des patients à haut risque nécessitant des soins spécifiques (de nature similaire à un système d'aide au recrutement ou d'attribution d'un crédit). L'un des paramètres utilisés par ce système étant le montant des dépenses de santé, il recommandait ces soins pour des patients blancs beaucoup plus souvent que pour des patients noirs[11].

[9] https://www.youtube.com/watch?v=9rDo1QxI260

[10] https://www.loria.fr/fr/evm-un-algorithme-predictif-de-linsuffisance-cardiaque-cree-a-nancy-avec-lia/

[11] http://www.slate.fr/story/183384/systeme-sante-americain-algorithme-raciste

À côté de nombreux succès, il faudrait pour le moins supprimer *personnalisée*, pour des systèmes basés sur l'apprentissage automatique, dans les objectifs de la médecine 4P – prédictive, préventive, personnalisée et participative –, car si les statistiques sont vraies en général (en santé publique, les méthodes bayésiennes fournissent des preuves statistiques de l'influence de certains facteurs sur certaines maladies) elles sont fausses en particulier (parce que dans ces méthodes chaque individu est un point isolé dans un espace de grande dimension), médecine personnalisée et médecine prédictive sont (presque) antinomiques[12] car une vérité statistique est vraie en général (en épidémiologie) mais peut être fausse en particulier (« mon voisin qui fume comme un pompier a 95 ans et se porte très bien ! »).

Recrutement

Face aux énormes quantités de CV reçus de nombreux services de ressources humaines utilisent des systèmes plus ou moins automatiques d'aide au recrutement. Les systèmes fonctionnent généralement de la manière suivante : ils analysent les CV et fournissent une note, seuls les CV ayant passé un certain seuil sont lus, dans le sens lus par des humains. Le système développé puis abandonné par Amazon a été souvent commenté. En effet, ce système de recrutement avait été entraîné sur des données concernant majoritairement des hommes, le résultat fut de sous-noter les femmes et de sur-noter les hommes « leur logiciel de recrutement n'aime pas les femmes » a pu titrer Reuters[13]. « L'entreprise a tenté de le modifier pour le rendre neutre, mais a finalement estimé qu'elle était dans l'impossibilité de garantir que le logiciel n'apprendrait pas d'autres façons discriminatoires de trier les candidats et a mis fin au projet. »[14]

Reconnaissance de genre

Le même type de biais a été mis en évidence par *Joy Buolamwini,*[15], une chercheuse du MIT, qui a démontré que les systèmes de reconnaissance du genre d'une personne à partir d'une photo faisaient beaucoup plus d'erreurs pour le classement des femmes *à la peau sombre* que pour celui des hommes blancs. Par exemple, le système de Microsoft classifiait incorrectement 1% d'hommes blancs et 35% de femmes à la peau sombre. Pourquoi ? Parce que leurs données d'apprentissage comprenaient beaucoup plus d'hommes blancs que de femmes noires.

Police

PredPol est un système de police prédictive basé sur une méthode pour prévoir les répliques des tremblements de terre. Son objectif était de déterminer les secteurs où la police devait être prioritaire. Les données utilisées étaient celles issues des plaintes des victimes, sauf pour les homicides, et pas celles des arrestations (sans doute pour ne pas être politiquement incorrect). Or, les enquêtes concernant les personnes s'affirmant victimes « montrent que la distribution des plaintes n'est pas homogène dans la population car certaines victimes pensent que la police ne peut rien faire pour régler leurs problèmes ou qu'il ne vaut pas la peine de déposer plainte. Le fait que les victimes ne recourent pas à la police s'explique par leur position sociale, leurs expériences passées avec la police, leur lieu de résidence et leur propension à agir dans l'intérêt de la vie du quartier. Les non-signalements sont des phénomènes sociaux en tant que tels, qui

[12] https://theconversation.com/medecine-police-justice-lintelligence-artificielle-a-de-reelles-limites-170754

[13] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[14] https://www.lebigdata.fr/amazon-abandonne-ia-misogyne

[15] https://fr.wikipedia.org/wiki/Joy_Buolamwini

échappent complètement à l'apprentissage statistiques par les données enregistrées par la police. » Ainsi PredPol qui conduit à ce que comme l'écrit Cathy O'Neil « Les innocents entourés de criminels se font malmener, tandis que les criminels entourés de bons citoyens respectueux des lois passent au travers ; Et compte tenu de la forte corrélation entre les situations de pauvreté et le signalement de délits, ce sont les plus modestes qui continuent d'être pris au piège de ces filets numériques » a été abandonnée par la police de Los Angeles[16].

Refus de demande de prêt

Les refus de demande de prêt, basé sur des systèmes intégrant des méthodes d'apprentissages, est un autre exemple de décision tellement fréquent que les médias s'en sont emparés[17]. Un chercheur en IA, mondialement connu, ayant vu une demande de prêt immobilier refusée alors qu'il avait des revenus suffisants et pérennes a réussi à obtenir le code du programme et il a découvert que c'était parce que le précédent propriétaire ainsi que de nombreux voisins de la maison qu'il voulait acheter avait des dettes !

Justice

Les finalités du système Datajust, basé sur de l'apprentissage à partir des données du ministère de la Justice, sont décrites ainsi par ce ministère[18] :

> « 1° La réalisation d'évaluations rétrospectives et prospectives des politiques publiques en matière de responsabilité civile ou administrative ;
> 2° L'élaboration d'un référentiel indicatif d'indemnisation des préjudices corporels ;
> 3° L'information des parties et l'aide à l'évaluation du montant de l'indemnisation à laquelle les victimes peuvent prétendre afin de favoriser un règlement amiable des litiges ;
> 4° L'information ou la documentation des juges appelés à statuer sur des demandes d'indemnisation des préjudices corporels. »

Malgré une requête déposée devant le Conseil d'État au nom de plusieurs associations, celui-ci a accepté qu'une période d'expérimentation de ce système se déroule jusqu'en mars 2022. Comme pour la médecine personnalisée, la justice ne peut être rendue automatiquement, elle est « personnalisée » dans son principe même si les juges peuvent être aidés par des systèmes automatiques, dans la mesure où ils ne se fient pas aux propositions du système, que ce soit pour gagner du temps, par pression de leur hiérarchie ou tout simplement à cause d'une trop grande confiance dans les résultats fournis par ces systèmes.

Pour terminer cette partie mentionnons que ces problèmes se posent dans d'autres domaines. Par exemple, de nombreux systèmes de traitement automatique des langues ont été construits en utilisant la plateforme GPT3[19] développée par OpenAI[20], qui utilise des milliards de variables, et permet, par exemple, après un mot de prévoir le mot suivant… mais avant de l'utiliser, comme dans tous les exemples précédents, il est nécessaire de s'interroger sur les qualités des données, en particulier les possibles biais dans les corpus utilisés pour l'apprentissage. S'il s'agit, par exemple, d'un système de

---

[16] Cathy O'Neil *Algorithme. La bombe à retardement* (préface de Cédric Villani. Les Arênes, 2018)

[17] https://theconversation.com/discrimination-et-ia-comment-limiter-les-risques-en-matiere-de-credit-bancaire-167008

[18] https://www.justice.fr/donnees-personnelles/datajust

[19] https://www.legifrance.gouv.fr/loda/id/JORFTEXT000041763205/

[20] https://openai.com/blog/gpt-3-apps/

traitement de la parole, les accents et les problèmes d'élocution des locuteurs sont à analyser.

Limites scientifiques

Comme tout programme informatique un logiciel intégrant des techniques d'IA est soumis à des limites intrinsèques concernant, entre autres, son exactitude et sa complexité. Les problèmes abordés avec des techniques d'IA sont généralement très compliqués (voire indécidables) et les solutions obtenues sont ainsi nécessairement imparfaites : elles sont incomplètes (ne concernent pas tous les cas) et souvent seulement approchées (ne sont pas optimales). De plus, s'il est généralement difficile de prouver (mathématiquement) qu'un programme fait bien ce qu'on voudrait qu'il fasse, dans le cas de l'IA il est également difficile de déterminer précisément son domaine de validité et sa marge d'erreur. On peut se convaincre facilement que c'est le cas dans un système de conduite autonome mais aussi dans les exemples donnés ci-dessus.

D'où l'importance fondamentale d'une tâche particulièrement difficile : faire une « bonne » évaluation du système qui concerne, elle aussi, les statistiques et les sciences humaines et sociales.

Adaptation de l'utilisateur au programme : de l'identité numérique à l'homme numérique

Dans la plupart des systèmes d'aide à la décision intégrant de l'IA, l'utilisateur, sauf s'il a de sérieuses connaissances scientifiques, un esprit critique développé et le sens de ses responsabilités personnelles pour pouvoir éventuellement contester ou s'opposer à un supérieur, risque de devenir un simple intermédiaire, de s'adapter au programme et pas l'inverse (beaucoup plus coûteux). Le décideur est amené à modifier son comportement, sans probablement s'en apercevoir, parce qu'il ne comprend pas le fonctionnement du programme, ne le comprenant pas, il se fie à ce qui ne devrait être qu'une proposition pour en faire une décision, et le système d'aide risque de se transformer ainsi, imperceptiblement, insidieusement, en un système automatique, or un homme est plus complexe qu'une centrale nucléaire.

Dans le cas d'un système grand public c'est M. ou Mme Tout-le-monde qui risque d'être amené à se conformer au système pour l'utiliser au mieux. Ceci n'est pas spécifiquement lié à l'IA, c'est le cas de tout outil informatique, et même de tout outil, y compris une scie ou un marteau. Mais une scie n'a pas de modèle de l'utilisateur. Les systèmes intégrant de l'IA et concernant des personnes, contiennent des représentations informatiques, implicites ou explicites, de ces personnes qui seront amenées pour utiliser efficacement ces applications à ressembler à ces représentations numériques. Les systèmes de recrutement nous contraignent à nous conformer au « bon employé » du système ; les systèmes de traitement des langues nous apprendre à parler ou écrire d'une certaine manière ; la voiture autonome ne nous apprendra pas à conduire correctement mais à comprendre comment l'automate conduit et au pire à conduire comme un programme ; le portefeuille d'identités numériques veut nous faciliter les accès à certains services en nous remplaçant par des QR codes etc.

Terminons par un dernier exemple, les systèmes de rencontres. Jessica Pidoux a étudié 22 tels systèmes dans sa thèse soutenue récemment à 'École Polytechnique Fédérale de Lausanne[21]. L'objectif d'un système de rencontres est d'apparier deux

[21] https://infoscience.epfl.ch/record/288400?ln=en

personnes au travers de l'interface dans laquelle elles essayent de se décrire et de décrire leurs désirs. Elles sont amenées à se conformer, pour augmenter leurs chances de recevoir des propositions de partenaires, à la façon dont elles imaginent que le logiciel fonctionne ce qui pourrait tendre à leur imposer un comportement amoureux ... De plus, certains systèmes apprenant à partir des actions des utilisateurs, ils peuvent, comme Tay, perpétuer ou amplifier des préjugés humains. « Tinder, par exemple, recommande des matchs basés sur un modèle patriarcal », explique Jessica Pidoux car « Le système apprend que certains hommes plus âgés préfèrent les profils de femmes plus jeunes avec un niveau d'éducation inférieur, mais l'algorithme pourrait alors suggérer le même modèle à d'autres utilisatrices de l'application. »

L'IA n'est jamais seule. Même dans les programmes de jeux ce n'est qu'un outil parmi d'autres outils informatiques, de plus des IAs sont non seulement présentes dans les téléphones, les tablettes et les ordinateurs reliés via internet mais elles sont aussi très souvent présentes dans les objets connectés, les IAs font partie de l'*Internet of Things* (IoT). Un objet connecté est un objet possédant la capacité d'échanger des données avec d'autres entités physiques ou numériques via internet. L'IoT est une extension du réseau internet à des objets du monde physique via des capteurs et des effecteurs. C'est un réseau de réseaux permettant de relier des entités informatiques et des objets physiques pour échanger, stocker et traiter automatiquement des données. Ces dix dernières années l'IoT a connu une progression fulgurante, le nombre d'objets connectés hors téléphones et ordinateurs de tout type (y compris les tablettes), présents dans tous les domaines, dépasse la moitié du nombre total d'objets connectés évalué à une centaine de milliards, et son chiffre d'affaires se compte en centaines de milliards d'euros ! Même si les données personnelles proprement dites sont protégées, le croisement des métadonnées permet d'obtenir de nombreuses informations (y compris parfois nominatives, car elles peuvent, dans certains cas, permettre de lever l'anonymat de données). Que ce soient à partir des métadonnées des e-mail expédiés ou reçus, (destinataire, date, objet, envoyé depuis un cybercafé ou de chez soi), des sms envoyés ou reçus (lieu, longueur), des photos (format, heure, lieu, appareil), de la navigation sur le web, des données fournies par des capteurs (consommation d'électricité ou d'eau) etc. les techniques permettent de connaitre des quantités d'information sur votre vie privée (sur vos positions politiques, religieuses, culturelles, sur votre santé, vos loisirs, vos achats, etc.).

Cette évolution va démultiplier les applications de l'IA en amplifiant bénéfices et risques. Les systèmes d'aide à la décision prolifèreront et comme ils incitent les utilisateurs à se comporter comme un programme il ne faudrait pas cela transforme leur irréductible singularité en un ensemble fini de symboles régi par des procédures de calcul !

Conclusion

Après l'Homme à l'image de Dieu qui fut un progrès considérable pour l'humanité, que faire pour empêcher qu'il soit transformé en un Homme numérique et l'ordinateur, ou l'IA, en un nouveau veau d'or ?

Pour éviter une catastrophe, pour éviter que la fuite en avant impulsée par les GAFAM et autres entreprises de la Tech conduise à l'émergence dans notre société à des fractures bien plus dangereuses qu'une « fracture numérique », pour éviter un effondrement de nos valeurs humanistes et en particulier la disparition du libre exercice de notre faculté de jugement et d'action, les quelques mesures suivantes me semblent nécessaires.

Formation

Dans le domaine de l'aide à la décision, mais plus généralement dans toute application informatique, en plus évidemment de connaissances en informatique, des connaissances en mathématique ainsi que la maitrise de la méthode expérimentale (particulièrement importantes dans le cas de systèmes prédictifs) sont nécessaires.

Toute application informatique à des conséquences sur les individus et la société donc une formation en sciences sociales et humaines est nécessaire. Ces conséquences pouvant être dramatiques l'Éthique doit avoir une place importante dans les processus de décision et dans la formation (ce que nous développons plus loin).

Naturellement, il ne s'agit que tous aient les mêmes connaissances et compétences, elles sont fonction de l'implication des personnes dans de tels projets. Cependant, tous, y compris les décideurs politiques (car il n'y a pas que le commun des mortels qui prend parfois l'IA pour une voyante ou une cartomancienne) devraient avoir les connaissances de base afin de pouvoir participer à un débat démocratique concernant les objectifs et les moyens de les atteindre, ainsi que l'évolution de ces domaines.

Débat démocratique

Les grandes entreprises présentent leurs projets comme un progrès et comme une évolution technologique inévitable. Elles sont puissantes et avancent rapidement. Pour ne pas être submergé par leurs objectifs, pour décider de la société dans laquelle nous voulons vivre, éducation et débat démocratique devraient être organisés par une structure neutre, donc par l'État (et pas par les réseaux sociaux ni par Marc Zuckerberg ou Elon Musk qui sont juges et parties).

Un tel débat démocratique devrait orienter l'élaboration de lois et de règlements régulant ces divers domaines, ainsi que des normes technologiques.

Pour ne donner qu'un exemple, il faudrait imposer la possibilité de se déconnecter facilement de l'IoT ! (Sa dénomination comme *Internet du Tout* fait frémir.) Chacun devrait pouvoir – si ce n'est être comme dans une chambre anéchoïde ce qui semble techniquement difficile – se déconnecter de tout réseau quand il le souhaite. Ainsi, toute batterie de téléphone, d'ordinateur etc. aussi (souvenons-nous de Steve Jobs ou Bill Gates mettant un scotch sur leur caméra même lorsqu'elle n'était pas activée …) de même pour les puces RFID, y compris passives.

Éthique

Il y a tellement d'applications inquiétantes que de nombreux organismes, dans les pays démocratiques, se sont dotés de commission d'éthique concernant l'IA, et 193 pays ont accepté récemment une résolution de l'UNESCO sur ce sujet[22]. En France l'alliance Allistene, qui regroupe CEA, CNRS, CPU, INRIA, Mines-Télécom, Écoles d'ingénieurs, a créé la CERNA[23] la Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique. Tout programme, tout système informatique, rappelle la CERNA dans un rapport sur l'*Éthique de la recherche en apprentissage machine* devrait respecter de nombreuses propriétés :

- *loyauté* (un système informatique est loyal s'il se comporte comme ses concepteurs le déclarent),
- *équité* (l'équité d'un système informatique consiste en un traitement juste et équitable des usagers),

[22] https://news.un.org/fr/story/2021/11/1109412
[23] https://www.allistene.fr/cerna/

- *transparence* (la transparence d'un système signifie qu'un utilisateur peut vérifier son comportement),
- *traçabilité* (la mise à disposition d'informations sur ses actions suffisamment détaillées pour qu'il soit possible après coup de suivre ses actions),
- *explicabilité* (le fonctionnement d'un système doit être compris par un utilisateur),
- *responsabilité* (le donneur d'ordre ou le concepteur étant responsable si le système est mal conçu, l'utilisateur étant responsable s'il a mal utilisé le système),
- *conformité* (un système numérique doit être conforme à son cahier des charges ce qui signifie que le système est conçu pour effectuer les tâches spécifiées en respectant les contraintes explicitées dans ce cahier).

Toutes ces propriétés devraient être vérifiées *avant* que le système soit utilisé, en analysant son code et ses données. Cependant, ces diverses propriétés peuvent être difficiles voire impossibles à réaliser dans des systèmes utilisant des techniques d'IA (en particulier l'explicabilité pour les systèmes d'apprentissage numérique). De plus, il y a une contradiction entre la rapidité des innovations techniques et la lenteur de leur évaluation qui nécessite : la compréhension des conséquences, le contrôle non seulement de la conformité d'un programme aux lois mais aussi des propriétés que devraient avoir tout système informatique, ainsi que des procédures démocratiques, nécessairement lentes, afin que les citoyens puissent décider librement de leur destin.

L'urgence, en particulier de conquérir des marchés, dicte souvent sa loi, de plus, la restriction, voire la suppression, de crédits dans de nombreux secteurs, a conduit au remplacement de personnes par des programmes.

Certains ont poussé des cris d'orfraie contre les robots tueurs, ils avaient raison. Le contrôle par des personnes d'armes automatiques n'est pas toujours faisable, par exemple le chemin suivi par un drone dans un essaim de drones menant une attaque ne peut pas être contrôlé par des humains, et les conséquences peuvent être dramatiques, mais il n'y a pas que les robots tueurs… voici un extrait de ce même rapport de la CERNA :

> « De façon caricaturale, un véhicule autonome qui se trouverait à choisir entre sacrifier son jeune passager, ou deux enfants imprudents, ou un vieux cycliste en règle, pourrait être programmé selon une éthique de la vertu d'Aristote – ici l'abnégation – s'il sacrifie le passager, selon une éthique déontique de respect du code de la route s'il sacrifie les enfants, et selon une éthique conséquentialiste s'il sacrifie le cycliste – ici en minimisant le nombre d'années de vie perdues. Le propos n'est pas ici de traiter de telles questions qui relèvent de la société toute entière… »

On peut se demander s'il est éthiquement responsable de continuer à essayer de construire des voitures autonomes, et plus généralement des véhicules autonomes ne circulant pas sur des voies spécialement aménagées. En effet, s'il y a un obstacle Météor, Orlyval, un tramway ou un train n'ont pas à choisir, ils freinent pour éviter le choc ! Il semble inadmissible qu'on puisse demander à un algorithme de choisir les personnes à sacrifier, et donc demander à un informaticien de programmer de tels choix !

Le livre de Cathy O'Neil déjà cité – qui ne demande aucune compétence technique pour sa lecture – contient de nombreuses études bien documentées sur ce sujet et Cédric Villani termine ainsi la préface de ce livre : « Elle [Cathy O'Neil] est bien décidée à ne pas baisser les bras, et à faire porter sa voix autant qu'il le faut pour que nous puissions conserver notre humanité. Écoutez-la attentivement. »

Pour éviter les dangers inhérents à de nombreuses applications de l'IA, et plus généralement de l'informatique, une formation à l'éthique des étudiants, enseignants, ingénieurs, concepteurs et utilisateurs serait nécessaire.

Risques

Si de telles mesures n'étaient pas prises la méfiance vis-à-vis de la science et de toute rationalité s'amplifierait avec toutes ses conséquences : impossibilité d'un débat public rationnel conduisant à une augmentation des fractures de notre société, sabotages (par exemple un sondage de septembre 2020[24] nous apprend que si moins de 20 % des personnes interrogées se disaient favorables à la destruction des antennes-relais 5G, 48% se disaient favorables à la suspension du déploiement de la 5G, un débat démocratique aurait sans doute limité les nombreuses destructions d'antennes 5G ou de matériels d'Enedis), refus des vaccins etc.

Pour terminer, je voudrais revenir un instant à l'Intelligence Artificielle en tant que science en rappelant la première phrase de Turing dans son article fondateur de l'IA[25], dans lequel il propose de considérer la question *"Can machines think ?"* et dont l'une des dernières phrases est : « Nous pouvons espérer que les machines finiront par concurrencer les hommes dans tous les domaines purement intellectuels (*purely intellectual fields)*. »

Donc il n'est pas question de l'homme dans sa totalité, ni même de ses *compétences* intellectuelles mais seulement de ses *performances* dans les « *purely intellectual fields »*. Pour simplifier Turing proposait deux voies. La première était celle des jeux, et le succès d'AlphaGo le programme qui a battu Lee Sedol, 18 fois champion du monde, a perturbé le sens critique de certains, ce programme, répétons-le, est incapable d'expliquer pourquoi il a gagné, de commenter une partie, et est encore plus incapable d'apprendre à un humain à jouer au Go ! La deuxième voie était celle de la construction d'un système ayant les capacités humaines concernant le langage c'est-à-dire capable d'apprendre n'importe quelle langue et de communiquer et sur cette voie, on n'en est qu'au début !

L'Intelligence artificielle en tant que domaine scientifique ne devrait pas inquiéter, au contraire elle permet de mieux comprendre certaines de nos compétences et la quête d'une IA générale ou forte non liée aux compétences humaines, est un domaine de recherche passionnant comme tous les domaines concernés par cette conférence et par tant d'autres. Mais il faudrait que la société se donne les moyens pour contrôler démocratiquement et sérieusement ses applications dont les conséquences sur l'humanité sont capables du meilleur comme du pire.

[24] https://www.ifop.com/publication/les-francais-et-les-habitants-des-grandes-villes-soutiennent-ils-le-deploiement-de-la-5g-en-france/

[25] *Computing Machinery and Intelligence.* Mind, 1950