Intelligence artificielle, mythes et réalités

B. Fallery, document de travail, Equipe de recherche MRM-SI, Montpellier 2019
(tous les liens cités dans le texte sont actifs)

1. Les processus de l'IA-informatique : reconnaissance artificielle et logique artificielle

En informatique, l'IA a pour objet de décomposer des fonctions cognitives pour simuler des comportements humains dans ses différentes activités : perception-acquisition, mémoire-apprentissage, raisonnement-pensée, expression-communication et exécution-décision. Il s'agit bien de simulation, de reproduire des résultats et non des processus humains. Mais derrière ce programme général il faut distinguer l'approche de l'IA *numérique* (qui s'appuie sur les progrès de la statistique et des bases de données massives) et l'approche de l'IA *symbolique* (qui s'appuie sur les progrès de la logique formelle et de la représentation de connaissances).

1.1 L'IA numérique : la reconnaissance artificielle

Depuis une quinzaine d'années c'est l'approche de l'IA *numérique* (ou connexionniste) qui domine la discipline (*MIT Technology Review* 2019), par une conjonction entre trois types d'évolutions :

- des évolutions proprement informatiques : sur la gestion des données (bases de données massives et distribuées, *NoSQL, Hadoop, Cloud computing*...) et sur la puissance de calcul (calcul parallèle sur des grappes de serveurs *HPC,* coût et miniaturisation des nouveaux capteurs et des processeurs graphiques) ;
- des évolutions en statistiques sur la classification automatique, qui ont permis la fouille de données *:* processus de décision markoviens, algorithmes des k plus proches voisins, régressions, forêts d'arbres de décision, partitionnement de données similaires, calculs de singularités...
- et aujourd'hui des progrès dans l'utilisation les réseaux de neurones formels : le Perceptron, première conceptualisation d'un neurone formel apprenant, avait été proposé par F. Rosenblatt en 1957 (Cours *Youtube* 2010) mais c'est en 2019 que les travaux sur les architectures de réseaux de neurones ont valu le prix Turing à Y. LeCun , Y. Bengio et G. Hinton (LeCun 2016)

L'entrainement d'un réseau de neurones formels se fait uniquement « par expérience », soit en apprentissage supervisé sur des données déjà labellisées, soit en apprentissage par renforcement suivant des valeurs de récompenses. La masse considérable de données nécessaires devenant très couteuse en puissance de calcul, les recherches portent autant sur des processeurs spécialisés que sur de nouvelles architectures de matrices. Certaines architectures complexes amènent alors Y. Bengio (2019) à parler de « la révolution de l'apprentissage profond » (c'est la structure des réseaux qui est profonde) :

- les réseaux convolutifs ont par exemple la particularité d'utiliser une hiérarchie de matrices-filtres (pour ne traiter qu'une portion de l'information) qui génèrent des matrices-maps ou patterns (LeCun 2016) : ces réseaux sont alors capables distinguer des formes avec des niveaux d'abstraction successifs, pour analyser pas à pas une image ou un son ;
- les réseaux récurrents ont par exemple la particularité d'avoir une structure qui n'est pas uniquement linéaire, mais qui contient aussi des connexions sous forme d'arcs ou cycles de retro-action : en pouvant alors mémoriser et réinjecter des entrées passées, ces réseaux permettent l'analyse ou la construction de séries temporelles (les phonèmes pour la voix, les mots pour le langage, les notes pour une partition...).
- les réseaux génératifs adverses GAN ont par exemple la particularité de générer des formes et de

les soumettre à un réseau discriminateur qui peut rejeter celles qui ne lui apparaissent pas réelles : le processus d'apprentissage s'arrête quand le Générateur est capable de créer des formes qui trompent le Discriminateur à tous les coups (Wintics 2018).

Par la « simple » perception-analyse d'une masse de situations ponctuelles dans un domaine particulier, les réseaux de neurones formels infèrent notamment des fonctions de prédiction *par induction* très intégrées (estimations boursières, profils de clients, prédiction des fraudes, analyse financière, assistants virtuels Chatbots, systèmes de recommandation, tarification discriminante en temps réel, conduite autonome…), mais ils ne fournissent aucune règles ou justifications *pour l'interprétation*, puisque les résultats obtenus en sortie ne sont pas explicables : quand le réseau de matrices devient « profond » (*Deep learning :* jusqu'à 150 matrices dans le modèle Resnet, *Microsoft Research* 2015), les millions de modifications automatiques du potentiel de tous les neurones ne sont plus *traçables.*
Cette approche numérique ou connexionniste de l'IA connait des succès spectaculaires dans les domaines où on peut s'appuyer sur un *entraînement* suffisamment rapide d'un réseau de neurones et sur une masse considérable d'exemples : vision et reconnaissance d'images, reconnaissance faciale, reconnaissance de la voix et des émotions, reconnaissance du langage naturel et traduction, reconnaissance des perturbations génétiques... On voit ainsi apparaître la formule : IA = *Big data* + *Machine learning* (ce dernier représentant en effet 89% des 55.000 brevets déposés en IA en 2017, *OMPI*).

1.2 L'IA symbolique : la logique artificielle

C'est pourtant l'approche de l'IA *symbolique* (ou logique) qui avait fondé la discipline (Ezratty 2018), d'une part en considérant le raisonnement comme une manipulation de symboles s'appuyant sur la logique formelle (logique des propositions vraies ou fausses, logique des prédicats sur des variables quantifiables, logique floue avec des degrés de vérité...) et d'autre part en rendant possible une représentation du monde sous la forme d'ontologies formelles :

- il ne s'agit pas ici de reconnaître des formes et d'extraire des relations prédictives, il s'agit de reproduire un processus mental peu structuré en se basant sur des structures informatiques différentes des bases de données classiques : des ontologies formelles d'un domaine de connaissances, des réseaux et graphes conceptuels, des *frames*, des bases de règles, des systèmes multi-agents... Il s'agit donc de résolution de problèmes en exploitant des connaissances existantes dans des domaines déjà théorisés : démonstrateurs de théorèmes, moteurs de recherche sémantiques, parcours des arbres de décision, extraction de l'expertise... et un des intérêts majeurs pour l'aide à la décision est de pouvoir fournir ici une *explication* à la solution informatique proposée pour résoudre le problème : par exemple par la trace des règles utilisées par un moteur d'inférence dans un système expert ou par l'enchainement des échanges entre entités dans un système multi-agents (Ferber 1995) ;
- d'une part cette approche symbolique ou logique de l'IA n'est certes pas adaptée à la fonction de perception-reconnaissance et d'autre part les connaissances restent difficiles à représenter et à conceptualiser : elles sont à la fois tacites et explicites, et elles nécessitent des codages et des mises à jour qui sont encore fastidieux. Mais de nombreux chercheurs plaident néanmoins aujourd'hui pour son renouveau et son association avec l'approche numérique, dans un contexte où le *Machine Learning* uniquement statistique (toujours plus de données, toujours plus de processeurs) semble commencer à montrer les limites d'une « intelligence superficielle » qui ne peut pas intégrer un minimum de sens commun sous la forme de savoirs a priori (règles, modèles prédictifs, inférences...), *InternetActu* 2017, *L'UsineNouvelle* 2018. On se souvient par exemple de l'embrasement général après la victoire au jeu de Go (*Medium* 2017) d'AlphaGo Zero de DeepMind contre Lee Sedol... mais la machine a bizarrement pris sa retraite (*Znet* 2017) et on a appris en 2019 que cette IA de DeepMind, n'utilisant toujours

que *l'apprentissage par renforcement statistique*, s'est révélée incapable de passer un contrôle de mathématiques de niveau lycée pour ne pas savoir mettre un problème sous forme d'équation (*Le Figaro* 2019).

2. Les processus de l'intelligence humaine : créativité et esprit critique

En se limitant ici à la question de savoir si les processus de l'intelligence Humaine devraient craindre la concurrence des puissants processus de l'intelligence Informatique, deux points apparaissent essentiels : la créativité et l'esprit critique.

2.1 La créativité, comme différence entre « savoir reconnaître » et « créer une connaissance »

Pour savoir si une machine fait montre d'intelligence, on connait le succès de la proposition d'Alan Turing (1950) : un test à l'aveugle pour attribuer un dialogue à une machine ou à un humain, test donc uniquement basé sur la performance d'une imitation ressentie… et évitant ainsi justement d'avoir à définir l'intelligence. Même si ce « jeu de l'imitation » basé sur le langage s'est révélé être une incitation considérable pour le développement de l'IA, il a aussi fait oublier le fait qu'un même résultat intelligent et inventif (ce que peut faire une machine et de mieux en mieux : écrire une partition musicale, dessiner un tableau, composer un roman, imiter une voix, créer une page Wikipedia, tenir une conversation en reconnaissant des émotions, créer de l'empathie avec un chatbot..) n'implique pas des *processus identiques* pour l'obtenir :

- on a vu qu'en IA un système apprend et raisonne, soit en manipulant logiquement des symboles (dès 1963, c'était déjà la proposition de A. Newell et H. Simon), soit en manipulant numériquement des vecteurs et des matrices (c'est aujourd'hui la proposition de l'IA numérique). Un ordinateur peut donc être *inventif* dans la mesure où le résultat obtenu n'est non seulement pas prévisible mais peut même être entièrement généré, comme c'est le cas par exemple pour des images, des textes ou des architectures avec la nouvelle approche des réseaux de neurones GAN Generative Adversarial Network (*Wintics* 2018). Un robot peut donc être *autonome* dans la mesure où la régulation se fait autant par rapport à son état interne que par rapport à son environnement, on parle d'autorégulation ou même d'auto-organisation non prévue par le programme de départ ;

– mais un être humain a non seulement conscience de lui-même (comme certains animaux ou même certains robots « zombies » qui réussissent le « test du miroir », *Interstices* 2016), mais il *a aussi* conscience de penser, et *surtout* il le fait dans un contexte social, psychologique et temporel :

- conscience de penser dans un contexte social, en intégrant non seulement ses propres expériences *mais aussi* des connaissances de tiers de confiance et des controverses dans l'espace public ;

- conscience de penser dans un contexte psychologique et corporel, en faisant interagir non seulement plusieurs représentations qui sont connues *mais aussi* des sentiments, des souvenirs et des introspections qui associent le réel et l'imaginaire ;

- conscience de penser un contexte temporel, selon diverses logiques et objectifs qui peuvent être non seulement subis *mais aussi* qui peuvent être choisis, initiés ou adaptés selon le déroulement des circonstances.

Si un algorithme peut certes faire des découvertes étonnantes en exploitant des bases de données massives, il s'agit essentiellement de reconnaître des régularités ou des singularités. Un exemple : c'est le séquençage de l'ADN des échantillons prélevés lors de l'expédition Tara Océans qui a révélé

plusieurs milliers d'espèces de plancton qui étaient inconnues et dont un tiers ne peuvent même pas être rattachées à un groupe répertorié (*Pour la Science* 2019).

La créativité quant à elle est en revanche toujours le fruit de l'interaction avec le contexte, « un processus de conception d'une solution jugée nouvelle, innovante et pertinente *en lien* au contexte précis de la situation-problème » (Romero et al. 2017). Mais le plus important est que ce contexte précis est bien plus qu'un « environnement », il est bien sûr socio-culturel et socio-psychologique, mais il est même socio-historique : la vérification des théories sur les ondes gravitationnelles ou sur le boson de Higgs s'est ainsi étalée sur un siècle (*La Recherche* 2016). Quant à l'étage supérieur, c'est-à-dire celui d'un changement de paradigme, d'un bouleversement de la perspective sur le monde par l'association de nouveaux concepts, on comprend alors que cela n'a aucun sens pour une intelligence artificielle, même si elle devient un jour à la fois numérique et symbolique. La cognition humaine ne se résume pas à un simple empirisme perceptif en additionnant des images, et pour reprendre une jolie formule de D. Cardon (2015) « il est encore temps de dire aux algorithmes que nous ne sommes pas la somme imprécise et incomplète de nos comportements ».

2.3 L'esprit critique, comme antidote à la connerie humaine

En 1983 dans son ouvrage Où en est la psychologie de l'enfant, R. Zazzo a pu se permettre de reformuler la question de l'intelligence sous la forme « *Mais qu'est-ce que la connerie, madame?* » (in Beaumatin 2008), car pour lui « le contraire de la connerie, ce n'est pas la logique » (Zazzo 1983, page 47). En 2018, C. Hadji (2018) peut se permettre d'expliquer que « la tension qui oppose l'insuffisance de maîtrise à l'efficacité technique, et pour laquelle l'IA serait le pôle supérieur, ne se superpose pas avec une autre tension, celle qui oppose connerie et intelligence critique ».
« *Entre nous soit dit, bonnes gens/Pour reconnaître/Que l'on n'est pas intelligent/Il faudrait l'être* » (G. Brassens, dans Ceux qui ne pensent pas comme nous) : on voit que pour *prendre conscience* de ses insuffisances il faudrait déjà être capable d'esprit critique, c'est à dire savoir s'interroger sur la valeur et les conséquences de ses actes.
L'intelligence humaine est plurielle et suivant l'usage de nos capacités cognitives « on peut à la fois être con et intelligent » (Zazzo, 1983, page 48). C. Hadji (2018) donne comme exemple de « se mettre à quinze pour tabasser un lycéen sans défense » : la conscience critique « cette autre dimension de l'intelligence, serait en quelque sorte un antidote pour la connerie » (et alors les machines, ne pouvant exercer qu'une intelligence logico-mathématique, seraient définitivement à l'abri d'une connerie artificielle).

Le véritable danger pour l'intelligence Humaine ne serait alors pas aujourd'hui une concurrence déloyale de l'intelligence Informatique, mais au contraire celui d'un abandon de notre esprit critique et de tout libre arbitre (Scaruffi 2016) : non seulement l'homme risque la bêtise en se déchargeant de plus en plus de nombreuses fonctions intellectuelles comme la mémoire, le calcul, la lecture... (Desmurget 2019), mais il risque surtout la connerie, l'absence d'esprit critique, en imitant l'intelligence numérique ou logique de l'IA. Car finalement, ne devrait-on pas inverser la question du fameux jeu de l'imitation de Turing et se poser cette question nouvelle : aujourd'hui, entre un humain et une machine, Qui imite Qui ? (Marosan, 2019). Ne sommes-nous pas en train d'intérioriser, sous la pression d'une optimisation ressentie comme nécessaire, une bonne partie des processus de l'intelligence artificielle : adapter son langage pour se faire mieux comprendre par son assistant vocal, adapter ses choix en fonction des recommandations Netflix ou Amazon, adapter ses mouvements pour améliorer un score de santé sur sa montre connectée, adapter sa lecture des actualités en fonction de son fil d'info Facebook, adapter ses conversations au rythme du multi-tâche imposé par ce téléphone mobile omniprésent... les êtres humains, eux, savent se faire robots.

Au final, et de la même manière que personne n'a eu l'idée d'appeler un avion un « oiseau

artificiel », on dira que l'intelligence artificielle fonctionne suivant d'autres processus et qu'elle n'a besoin ni de créativité ni d'esprit critique pour être intelligente à sa manière. A l'avenir, l'IA sera surement composée de nombreuses IA différentes, dans des domaines spécifiques et avec des capacités spécifiques en perception, apprentissage, raisonnement, communication, exécution. Ces capacités sont et seront souvent supérieures aux capacités humaines dans des domaine particuliers, mais cela ne mène pas à une intelligence reposant sur la créativité et l'esprit critique. Pour reprendre notre parallèle, personne ne se demande quand l'intelligence d'un avion aura dépassé l'intelligence d'un oiseau.

3. Les grands mythes de l'IA-fiction

3.1 Le mythe de la singularité, nouvelle version de Prométhée, de Pygmalion, du Golem

Prométhée a volé le feu sacré de l'Olympe pour en faire don aux humains qui n'avaient rien, ni la force des lions ni les ailes des oiseaux. Feu divin, feu connaissance, technologie primordiale mais dangereuse. Prométhée le titan sera puni par Zeus, son foie sera dévoré chaque jour par l'Aigle du Caucase, car son énorme ambition pourrait amener les humains à se surpasser. Morale de la fable : le rêve des humains qui voudraient que la technologie les fasse entrer dans le monde des dieux ne date pas d'hier.
Pygmalion est tellement fasciné par la perfection de cette statue qu'il a longtemps modelée, qu'il obtient d'Aphrodite qu'elle lui donne la vie. Il en tombe amoureux et épouse cette femme qu'il a créée de ses mains, devenue Galatée. Morale de la fable : le rêve des humains fascinés par la frontière entre eux et les objets qu'ils créent ne date pas d'hier.
Le Golem est apparu au XVIe siècle dans une communauté juive de Prague qui vivait dans la peur des calomnies. Pour se protéger et sur recommandation divine, le sage Yehoudah Loew sculpta alors un homme dans la glaise de la rivière avant de lui insuffler la vie selon un rituel secret. Mais sa création finit par se retourner contre ses maîtres et le sage dût lui ôter la vie. Morale de la fable : la fascination des humains face au pouvoir à la fois bénéfique et maléfique d'objets créés à leur image ne date pas d'hier.

Il y a fort à parier que Dmitry Itskov, jeune entrepreneur russe qui a lancé son « Initiative 2045 » a du être fasciné par ces mythes ancestraux sur la maitrise de la vie. Pourquoi 2045 ? Parce que les calculs montrent que c'est à cette date que le « point de singularité technologique » sera atteint. Sur le site http://2045.com/ on peut donc voir toutes les étapes qui vont nous amener en 2045 à l'apparition de l'avatar de type C, celui dont l'intelligence commencera à dépasser l'intelligence humaine puisque nous pourrons y télécharger notre propre esprit. Le site affiche plus de 47.000 abonnés et non des moindres puisqu'il compte en autres le Dalai-lama et K. Kurzweil, un des nombreux directeurs du développement chez Google.
La « démonstration » du point de singularité est en effet le cheval de bataille de K. Kurzweil (Conférence TED, la *Singularity University)*. En mathématique un point de singularité est un point critique pour une fonction (par exemple, 1/x quand x tend vers zéro), mais ici il ne s'agit que du point d'intersection de deux courbes: celle du développement de l'intelligence humaine, qui est croissante depuis la préhistoire mais pratiquement linéaire, et celle du développement de l'intelligence artificielle qui est exponentielle depuis 1950. Les deux courbes vont donc se croiser en 2045, l'argument étant que les capacités de l'IA vont suivre une loi identique à la fameuse « loi de Moore » sur le doublement tous les deux ans de la puissance des microprocesseurs (voir Cerebras le plus grand processeur d'IA, Ozeratty 2019). On passera ainsi de l'IA faible d'aujourd'hui (*Artificial Weak Intelligence*) à l'IA Forte de 2045 (*Artificial General Intelligence*), avant de voir s'épanouir bientôt la Super-IA (*Artificial Super Intelligence*).
La discussion de cette thèse, simple et spectaculaire, agite les médias en quête d'audience

puisqu'elle donne des frissons à ceux qui aiment se faire peur avec la technologie : de l'IAllégresse... jusqu'à IApocalypse. Même une église de l'IA est née en 2017, WayOfTheFuture. Mais l'hystérie prophétique de la Super-IA agite aussi une partie des experts : quelques grands scientifiques dont l'informatique n'est pas le domaine, tel fut le cas de Stephen Hawking, mais surtout des ingénieurs travaillant dans les GAFAMI qui se donnent le vertige d'appartenir aux géants du numérique qui pourraient bouleverser l'humanité.
L'analyse de J-G Ganascia (Ganascia 2017) sur ce mythe de la singularité permet de garder la tête un peu plus froide (Conférence Montpellier 2018).

3.2 Le mythe du transhumanisme, nouvelle version du Phénix, du Juif errant, de la fontaine de Jouvence

Le Phénix est un aigle gigantesque et magnifique de l'ancienne Egypte, mais il ne peut se reproduire. Alors quand il se sent trop vieux, il construit son nid et il y met le feu. Mais des cendres de ce bûcher surgit alors le Phénix nouveau.

Le Juif errant est un cordonnier de Jérusalem qui, s'étant moqué du Christ portant sa croix, a été condamné à l'immortalité : il vieillit jusqu'à l'âge de 100 ans, il tombe alors malade, mais quand il guérit il a à nouveau trente ans. Au XVe siècle le grand chancelier de Florence l'a rencontré sous les traits de Giovanni Votaddio, au XVIe siècle on le retrouve en Allemagne sous les traits d'Ahasvérus.

La fontaine de Jouvence est une source qui restaure la jeunesse, il suffit d'en boire quelques gorgées. Jupiter a transformé la nymphe Jouvence en fontaine, mais où est donc cette source ? Alexandre le Grand l'a cherchée en Asie, l'empereur chinois Qin Shi ne l'ayant pas trouvée non plus il s'est fait enterrer pour l'éternité avec son armée de terre cuite, des druides l'auraient utilisée dans la forêt de Brocéliandre, des conquistadors l'ont cherchée vers la Floride...

ll y a fort à parier que les nouveaux entrepreneurs des *Biotech* dans le secteur appelé maintenant NBIC (nanotechnologies, biotechnologies, informatique et sciences cognitives) ont du être fascinés par ces mythes ancestraux sur la maitrise de la mort. Citons Elon Musk (fondateur de SpaceX et de Telsa) qui vient de créer Neuralink pour développer des implants cérébraux connectés à des ordinateurs ; citons Larry Page (fondateur de Google) qui vient de créer Calico pour chercher les codes génétiques expliquant la longévité humaine et son hérédité ; en France et plus modestement citons Laurent Alexandre, ancien médecin devenu entrepreneur (fondateur de Doctissimo, propriétaire de DNAVision) mais aussi auteur prolifique sur les thèses du transhumanisme (« certains d'entre vous dans cette salle vivront mille ans », conclusion de sa conférence TED 2012).

Pour le transhumanisme, une autre forme d'intelligence artificielle va émerger : la reproduction *in silico* du fonctionnement du cerveau va se combiner avec les modifications aujourd'hui possibles de l'ADN, pour créer le cyborg, cet être humain hybride. Dans cet imaginaire, la technologie est sans limite : on passe ainsi des réalités sur l'homme augmenté ou sur la thérapie génique, pour aller vers l'abolition des maladies et de la vieillesse par la techno-médecine et les nano-robots... et donc bientôt vers l'abolition de la mort ; l'apparition d'une post-humanité. Les scientifiques du domaine parlent d'imposture (Tritsch et Mariani, *Pour la Science* 2018), ce qui n'empêche pas le sérieux Beijing Genomics Institute, par exemple, d'autoriser le séquençage de milliers d'ADN humains pour chercher à identifier tous les « gènes de l'intelligence » (*Sciences Humaines* 2019).