POUR UNE IA RESPONSABLE

PAR **DAVID CORTES** (97)

Il y a vingt ans, Pierre-Gilles de Gennes présentait ses travaux novateurs sur les cristaux liquides à travers la France et conseillait à son public, souvent réfractaire à l'étude de la chimie, de s'essayer à la lecture du livre *Le Système périodique* de Primo Levi, collection de récits, chacun inspiré d'un épisode autobiographique et d'un élément du tableau de Mendeleïev. Cette invitation à faire se croiser différents regards, et à éclairer les uns par les autres, vaut plus que jamais : convier littérature, sciences mathématiques, sociologie me semble bienvenu pour mieux saisir les enjeux de l'« IA ».

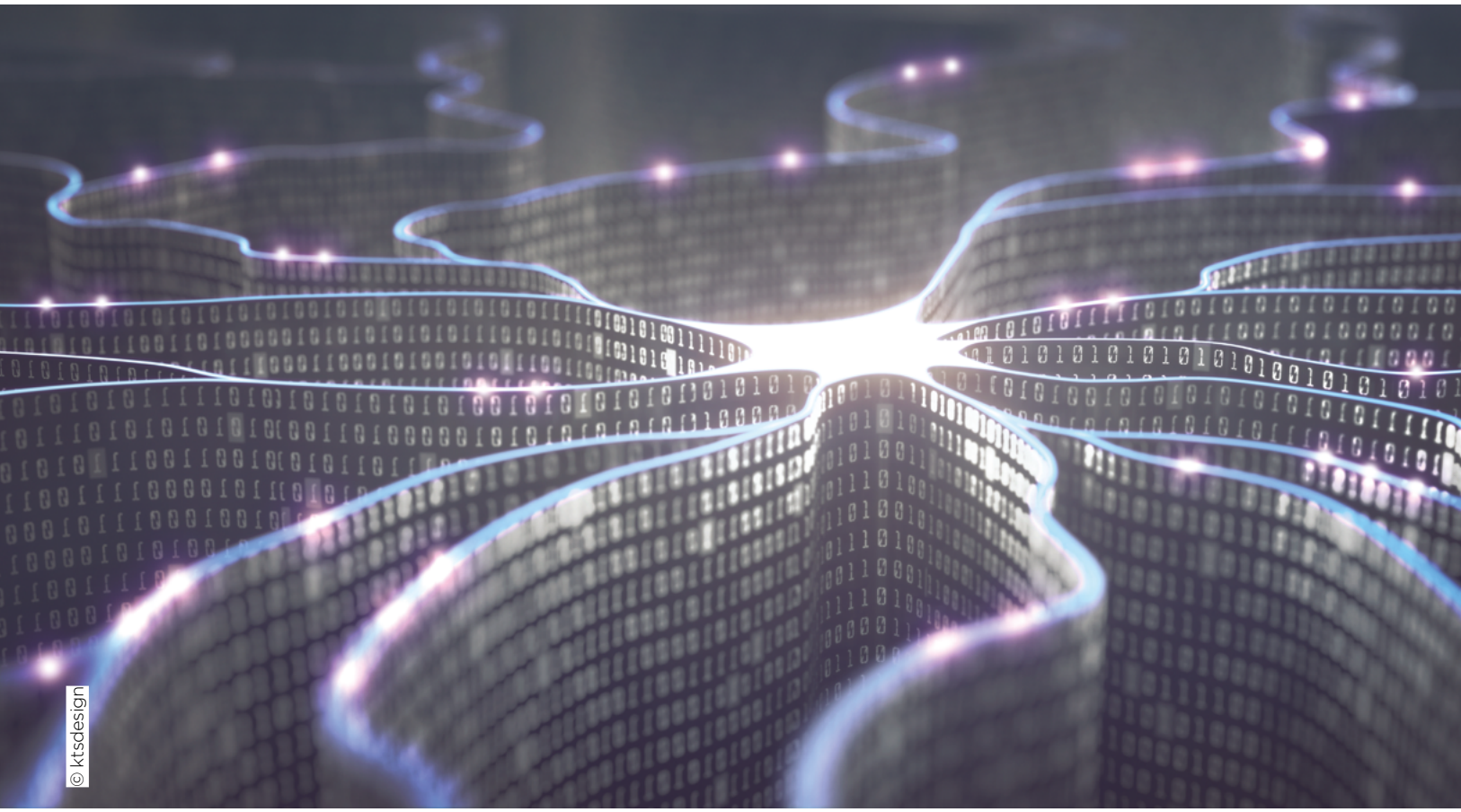
La recherche concernant l'« Intelligence artificielle (IA) » trouve son origine dans la conférence de Dartmouth en 1956, organisée principalement par John McCarthy et Marvin Minsky. Elle est le prolongement direct de mouvements plus anciens, tels que la cybernétique de Norbert Wiener développée dès 1947.

Dès 1956 : logique formelle et apprentissage machine

Les technologies présentées lors de cette conférence ont cependant fait date : dès 1956 l'algorithme de preuve formelle, permettant à des machines de démontrer certains résultats mathématiques ; et un embryon algorithmique de machines expertes au jeu des échecs. Puis en 1957 le perceptron de Frank Rosenblatt, élément constitutif des réseaux de neurones artificiels actuels. Ce dernier introduit en pionnier un concept majeur car, si les premiers algorithmes partaient d'une *approche déductive* (= partir de règles et prédire), ce dernier appelle une *approche inductive* (= partir des résultats et bâtir un système de prédiction) rendue possible par un système d'inspiration biomimétique. La machine doit donc « apprendre » mais, fait déroutant, sans même… comprendre, ni pouvoir expliciter son cheminement.
« L'IA » serait ainsi une nouvelle impulsion donnée à la fois aux *sciences cognitives* et à certains *systèmes informatiques* afin « d'émuler » les capacités mentales humaines. L'« IA » n'est donc pas une technologie, mais un champ de recherche qui vise en asymptote, lointaine, la recréation d'une véritable « intelligence humaine sur des substrats artificiels », variés. Les systèmes actuels, s'ils excellent sur certaines applications de perception, sont loin d'atteindre encore cet objectif.

Intelligence humaine ? l'IA actuelle est plutôt une « intuition artificielle »

L'intelligence humaine était vue par les informaticiens et mathématiciens de Darmouth dans un continuum de pensée entre *rationalisme* et *empirisme*. Stéphane Mallat rappelle dans les leçons au Collège de France que les approches de l'intelligence (*inter-legere*, « choisir entre ») évoluent sans chronologie depuis la pure spéculation abstraite jusqu'à l'apprentissage au plus près de l'expérience. L'intelligence humaine est à considérer non seulement comme innée, structurelle et purement logique, mais également comme largement construite, définie par un long processus d'apprentissages successifs nourri des corrélations observées lors de multiples expériences comparables.
Le terme d'intelligence est donc actuellement ambigu et souvent abusif. Ambigu, car en France, pays au cartésianisme encore triomphant, l'intelligence est vue principalement comme mathématique, déductive,



causale… beaucoup plus que dans des cultures, anglo-saxonnes par exemple, privilégiant l'empirisme, l'intuition. Abusif, car tout d'abord, comme le rappelle Yann Le Cun, le champ cognitif couvert par les « intelligences artificielles » actuelles reste extrêmement étroit : le plus puissant des supercalculateurs couplé aux algorithmes les plus innovants est très loin de savoir traiter « autant de problèmes que ne le fait le cerveau même d'un rat ».

Ensuite, parce que ces « intelligences » nouvelles apprennent des données, mais… sans nous donner en retour de règles explicites. Les systèmes de réseaux de neurones actuels sont constitués de gigantesques matrices de chiffres. Les « règles » apprises par ces systèmes sont distribuées dans l'ensemble des nœuds du réseau neuronal et ses millions de paramètres. Elles sont donc opaques par construction : de même que ce n'est pas en étudiant une image d'IRM, fût-elle prise à la granularité des neurones activés ou inhibés (~80 milliards dans un cerveau humain), que nous accéderons à l'instantané d'un raisonnement ou à la personnalité d'un patient.

Ces « intelligences » sont largement ainsi… des intuitions. Leur demander de nous expliquer leur fonctionnement supposerait d'elles à l'heure actuelle un effort d'introspection que leur principe de construction rend particulièrement difficile à réaliser.

L'IA, nouvel « or noir » : parachèvement d'une tendance biséculaire d'automatisation

Conséquence de l'engouement pour l'IA, sont apparues de très nombreuses études anxiogènes liées à l'automatisation : « Quelles professions sont les plus menacées, en fonction du secteur, ou de la plus ou moindre grande répétitivité des tâches, du pays, etc. Serai-je touché moi aussi ? » Si ces refrains lancinants sont en fait connus depuis plus de deux siècles, et les luddites de 1811 en Angleterre ou les canuts lyonnais, l'IA me semble toutefois porter en soi, conceptuellement, de quoi mener le phénomène d'automatisation à son terme.

En 1958, soit deux ans après la conférence de Dartmouth, Hannah Arendt proposait dans *Condition de l'homme moderne* un regard philosophico-historique, notamment sur le travail, qui fit date. Elle rappelle tout d'abord la hiérarchie des valeurs tacitement admise par l'Occident depuis Aristote, plaçant tout en haut les actions non nécessaires à la survie… action politique, philosophie, →

→ contemplation (*theoria*) et en second plan les tâches que l'on définirait actuellement comme celles du bas de la pyramide de Maslow. Les deux derniers siècles ont vu l'automatisation croissante de ces dernières (agriculture, commerce…) grâce à l'exploitation conjuguée des ressources naturelles : matières premières et énergies. L'IA se propose d'achever le mouvement et de nous débarrasser, également, à terme du souci des tâches cognitives.

Impact en retour sur l'intelligence humaine ?

Georges Bernanos (dans *La France contre les robots*, 1947, notamment) anticipe les effets du machinisme sur la « matière humaine » même et, par là, sur les organisations politiques et économiques. C'est une préoccupation proche des intuitions de Marshall McLuhan (développées dans *La Galaxie Gutenberg*, 1962), qui soutenait que le média lui-même influençait l'homme, et pas seulement le message qu'il véhiculait. Ainsi, par exemple, l'homme de tradition orale surdéveloppe une hypermnésie. Dit autrement : si l'apprentissage manuel s'amenuise par l'automatisation, puis si l'apprentissage cognitif est réduit par l'IA, quel pourrait être l'impact sur la formation de l'intelligence humaine ? quel sera « l'Homme de l'IA » ?
Tout d'abord une bonne nouvelle : les prophètes d'apocalypse faisant leur fonds de commerce d'une séparation entre élite maîtrisant et utilisant les IA et le reste de l'humanité ont certainement tort, pour deux raisons. D'une part, l'automatisation des tâches touchera à terme toute la population, quelle que soit la « noblesse » perçue des tâches qui lui incombaient auparavant (tertiaire inclus). D'autre part, même les plus experts en IA, leurs concepteurs, sont et seront toujours plus incapables de comprendre eux-mêmes les systèmes qu'ils auront créés, du fait de la complexité conjuguée des données et des algorithmes (ex. de l'ordre de 100 millions de paramètres pour les systèmes de reconnaissance d'images), et de l'empilement algorithmique (système de systèmes).

Intelligence artificielle oui, mais collective

Si les promesses de véritable intelligence artificielle *individuelle* renvoient à un horizon encore lointain, les formes d'intelligence artificielle *collective* portent d'ores et déjà le plus de fruits.
L'accroissement de précision d'IA individuelle *via* la prise en compte de beaucoup plus de facteurs prédictifs que ne peut le faire un cerveau humain contribue paradoxalement à certes renforcer la qualité des prédictions, mais dans le même temps réduit la variété des prévisions possibles.
Or cette réduction de diversité a également pour effet d'amoindrir la qualité des intelligences collectives. Émile Servan-Schreiber rappelle que par exemple le phénomène de sagesse des foules ne fonctionne que si la diversité des biais de chacun est suffisante pour que l'effet de moyenne sur le grand nombre permette une prédiction efficace. Les formes les plus efficaces d'intelligence collective actuelle (moteurs de recherche, graphes de causalité…) tirent leur force précisément de l'exploitation bien au-delà des capacités humaines individuelles, ou collectives, de cette diversité d'avis humains. Jusqu'aux algorithmes d'IA eux-mêmes, de types *Random Forest* ou *Boosting Trees* ou *Bootstrap,* qui génèrent un foisonnement d'algorithmes ou d'échantillons et en moyennent les prédictions, pour éviter le surapprentissage…
L'enjeu principal finalement ne serait-il pas, plus encore que dans l'éthique des IA, dans l'uniformisation croissante du savoir, des actions humaines ?

"L'IA actuelle est celle de l'apprentissage machine."

Pour une IA diversifiée et responsable

Si le terme d'intelligence artificielle est ancien, et souvent abusivement employé pour désigner les systèmes d'apprentissage statistique actuels, le retour en force de la notion d'apprentissage automatique a permis de réaliser des percées opérationnelles et conceptuelles majeures. Mais n'oublions pas pour autant, notamment dans nos cursus de formation, que ces systèmes sont le prolongement d'autres technologies parfois plus performantes sur certaines tâches, tels les systèmes experts ou les applications de recherche opérationnelle en général, et plus explicables. Encore insuffisamment soulignée, c'est toutefois l'intelligence artificielle collective qui porte les avancées majeures, mettant à profit les capacités informatiques afin de démultiplier les capacités d'analyse et de synthèse des savoirs humains. L'efficience sidérante des moteurs de recherche en témoigne.
Ces mécanismes d'apprentissage statistique reposant sur l'analogie posent principalement deux défis majeurs : leur opacité (par construction) et leur capacité normative à tendanciellement renforcer l'uniformisation globale. La question de l'éthique de ces IA appelle une vigilance nouvelle de la part de toutes les parties prenantes : citoyens, collaborateurs en entreprises, clients, comme régulateurs et législateurs.
Quelles que soient les caractéristiques des systèmes d'IA, un contrôle humain sera nécessaire. L'enjeu alors sera de garder la capacité de ces systèmes à toujours répondre de leurs décisions et s'expliquer : il est impératif d'œuvrer à une IA responsable. ×

Cet article, écrit en version plus développée pour *Variances.eu*, la revue des Ensae Alumni, a été publié le 13 février 2019. Nous le reprenons dans *La J & R* avec leur aimable autorisation.