# Quand l'intelligence artificielle théorisera les organisations

Philippe Baumard

# Quand l'intelligence artificielle théorisera les organisations

Philippe Baumard
*Revue française de gestion* 2019/8 (N° 285), pages 135 à 159

*RÉSUMÉ*

Cet article explore la possibilité qu'une intelligence machine puisse théoriser des organisations ; et qu'elle le fasse mieux qu'une intelligence humaine dans un proche futur. La plupart des modèles d'apprentissage des machines sont des processus statistiques automatisés qui sont à peine capables d'une induction formelle et ne génèrent pas de nouvelles théories, mais reconnaissent plutôt un ordre préexistant. Les théories humaines sont incarnées ; elles naissent d'un lien organique avec le monde, auquel les théoriciens ne peuvent échapper. Cet article envisage de surmonter cet obstacle pour accueillir une révolution théorique apportée par l'IA.

L'auteur tient à remercier William H. Starbuck, Wojciech Czakon et Michel Béra pour leurs commentaires et suggestions.

---

Ce jeudi 10 mars 2016, AlphaGo, le logiciel créé par les ingénieurs de Google Deepmind pour défier Lee Sedol, maître de go, fit un mouvement inattendu. AlphaGo jouait sa seconde partie dans ce qui deviendra sa victoire historique contre un des meilleurs maîtres de go humains. Alors que tous les regards sont concentrés sur deux espaces au sud du *goban*, AlphaGo place une pierre à l'intersection la plus incongrue qu'il soit : isolée, ouvrant un autre front ; un mouvement si étrange que tous les commentateurs de go se figèrent dans une absolue perplexité. À ce moment précis, Lee Sedol observa le *goban* attentivement, se leva et quitta la salle. Quand il revint, il était clair que ce mouvement étrange, singulier et inattendu était un trait de génie. Lee Sedol ne reprit jamais le dessus[1]. Ce mouvement fut largement commenté par les amateurs de go et les chercheurs en intelligence artificielle (McFarland, 2016). Mais il ne reçut que très peu d'attention de la part des théoriciens de l'organisation. Il constitue pourtant le signe distinctif d'un changement profond dans le futur de l'intelligence artificielle (ou IA), et comme nous le verrons dans cet article, annonce un changement profond dans l'activité humaine de la recherche en général, et de la recherche en management en particulier. AlphaGo avait-elle été programmée pour être surprenante afin de déstabiliser son adversaire humain, sans perdre ses chances de gagner ? Ou, au contraire, cette stratégie « créative » était-elle l'unique résultat de son algorithmique, sans supervision humaine ?

[1] https://www.washingtonpost.com/news/innovations/wp/2016/03/15/what-alphagos-sly-move-says-about-machine-creativity/?utm_term=.3ead89a16705

Nous parlons souvent, dans notre jargon scientifique, de « corpus de connaissances », comme si le caractère désincarné et autonome de la connaissance allait de soi. Il va de soi que l'on peut distinguer une connaissance à laquelle on a supprimé tout caractère particulier pour lui attribuer le statut de règle. Ainsi, *l'episteme* s'oppose dans la Grèce antique aux savoirs pratiques, techniques, à la sagesse ou à la sagacité particulière d'un individu (*techne*, *phronesis*, *eustochia*, etc.). Le savoir abstrait est avant tout un savoir d'*emprise* ; sa généralisation implique un statut désincarné : en retirant sa signature, le nom de son porteur, en le rendant distant, le savoir gagne en instrumentalité, en unilatéralité et en efficience. Nous avions déjà discuté comment les théories des organisations se sont construites dans une longue vassalité et servilité à ses donneurs d'ordre (Starbuck et Baumard, 2009). Nos théories servent toujours leurs pouvoirs contemporains, même, et parfois *surtout*, quand elles en font la critique. Aussi agaçant que cela puisse être, la théorie des organisations n'a jamais réussi à inverser le rôle du principal et de l'agent ; de l'institution et de ses membres ; du propriétaire et du manager ; du capital et de l'entrepreneur. À parler vrai, il n'existe aucune théorie de l'organisation qui ait réussi à démontrer une quelconque capacité d'auto-conception dans un fait organisé ; et cela même si la théorie du *self-designing organization*, introduite par Hedberg *et al.* (1976), fut tant de fois imitée.

Nous réinventons le *contenant* sans jamais révolutionner le *contenu*, parce que nous sommes profondément malléables et obéissants. Nos théorisations sont faites de tellement d'emprunts qu'à l'instar de *L'oreille cassée* d'Hergé, nous sommes condamnés à briser l'oreille de toute nouvelle théorisation pour la faire entrer dans le réel tolérable de cette soumission (voir Rosset, 1977). En d'autres termes, nous radotons. Nous sommes, en tant que communauté scientifique, celle qui excelle dans l'art de réinventer un corpus théorique qui dit toujours la même chose, remettant au goût du jour « un vieux vin dans une nouvelle bouteille » (Kimberly, 1981), mais n'en changeant jamais les mécanismes fondateurs ou la téléologie.

Le fait qu'une théorie soit « ancrée », ou qu'elle ait pu être éprouvée par le test de la réplication de ses résultats, a souvent peu d'importance dans cette téléologie. Prenons l'exemple des organisations matricielles. À l'origine, il s'agit de commandes de très grandes entreprises (DEC, HP, puis ABB) qui cherchent un modèle managérial pour pouvoir justifier des réorganisations impliquant de sérieuses réductions du nombre de cadres dirigeants dans la gestion de leurs marchés globaux. Aucune recherche n'est à l'origine du concept d'organisation matricielle. Un article

dans la *Harvard Business Review* rendit le concept populaire auprès de dirigeants luttant avec les difficultés de la naissante globalisation. Quelle autre fable avait le pouvoir de convaincre un cadre dirigeant d'assumer un double labeur de supervision de sa zone géographique et d'un produit sur le plan mondial ? Les exemples abondent. Nystrom et Starbuck (1984), dans leur classique « Façades organisationnelles », expliquent comment les diagrammes PERT (*Program Evaluation and Review Technique*) n'ont jamais réellement été un modèle utilisé pour gérer le programme du missile Polaris. « Étant donné que le PERT avait renforcé l'image de compétence de l'équipe de gestion, les comités d'examen externes se sont moins immiscés dans leurs affaires. Derrière cette façade, l'équipe n'a jamais réellement utilisé la technique pour gérer les échéanciers et les coûts. PERT n'a pas construit le Polaris, mais il a été extrêmement utile pour ceux qui ont construit le système d'armes de faire croire à de nombreuses personnes qu'il l'avait fait. » (Sapolsky, 1972, p. 125). Ainsi, le PERT et les innovations managériales associées ont fourni un « placage protecteur » (Sapolsky, 1972, p. 246).

Une théorisation de l'organisation ne peut jamais être complètement éprouvée par le test de la réalité. Les hommes aiment suivre des modèles, des « abrégés du vrai » aussi bien que des « abrégés du bon » comme aimait le dire Claude Riveline. Tant que ces abrégés fournissent une certaine communauté, tant qu'ils permettent de « faire sens » collectivement, comme l'exprimait Weick, d'assurer le maintien d'une subordination, ils confèrent une structure suffisante à l'action organisée. Peu importe qu'elle soit soutenue, ou pas, par un précepte scientifiquement établi. Peu importe que le subordonné soit plus ou moins dupe du manque de sincérité du modèle invoqué.

Cet « à peu près », cette tolérance à l'approximation, est ce qui motiva Jim March, Richard Cyert et Herbert Simon à poursuivre une théorisation du management qui accepte que les êtres humains puissent créer et maintenir des systèmes complexes, réaliser des prouesses techniques, en se gérant eux-mêmes avec des niveaux de « satisfaction minimale » (*satisficing*). L'organisation humaine possède une malléabilité cognitive et comportementale qui lui permet d'inventer *ex post* les théories qui expliquent ses actes, de créer des idéologies pour justifier des solutions, et de tolérer des écarts indéfendables entre ses théories professées et ses théories d'usage.

La victoire d'AlphaGo sur Lee Sedol peut très bien s'expliquer ainsi. AlphaGo a maximisé l'effet de perturbation sur son opposant en recherchant statistiquement les combinaisons les plus improbables, tout en maintenant une probabilité de succès

acceptable. Pourtant, si nous faisons l'effort de croire qu'une machine pourrait être réellement capable *d'invention*, sans passer par un mécanisme de leurre, cela pourrait-il constituer une lueur d'espoir pour les sciences de gestion, et peut-être, pour la recherche scientifique en général ? Car quelle hypothèse serait-elle plus optimiste que celle de pouvoir se débarrasser de cette épouvantable vassalité humaine ; du goût humain pour l'approximation, le plagiat et la répétition ; et des perversions institutionnelles de la publication scientifique ? C'est cet espoir qu'offre les timides avancées de l'intelligence artificielle contemporaine que cet article se propose d'explorer.

Cet article est organisé en trois parties. Dans un premier temps, nous étudions l'histoire de l'intelligence artificielle, depuis sa fondation au XIX$^{e}$ siècle, jusqu'à ses évolutions récentes, pour comprendre ce qu'une intelligence artificielle serait capable de faire en matière de théorisation… Ce qui nous amène, dans un second temps, à interroger l'acte de production scientifique afin d'identifier ce qui peut relever d'un acte humain, et ce qui peut faire l'objet d'une modélisation et d'un apprentissage autonome conduit par une machine. L'objectif est d'ici d'évaluer la faisabilité d'une substitution de l'homme par la machine pour produire une recherche.

Dans une troisième et dernière partie, nous proposons quatre modes d'exploration théorique qui sont déjà l'œuvre de machines, ou qui pourraient voir, dans le futur, une substitution complète de l'homme par la machine. Nous concluons cet article en partageant plusieurs interrogations sur l'avenir de la recherche en théorie des organisations, et son utilité, *homme ou machine*, pour les organisations et la société.

## I – À quoi rêvent les moutons électriques ?

Une machine ou un programme ne sont pas forcément condamnés à exécuter une série d'instructions pour reproduire ou imiter un comportement humain. Elles sont, pour la plupart, construites dans ce but. La première étape de notre réflexion est donc d'opposer, d'une part, la programmation dont l'objectif est de faire mieux et moins coûteux que la travail ou la cognition humaines ; et d'autre part, la programmation dont l'objectif est d'extraire une capacité heuristique humaine pour la transformer en capacité heuristique machine (Baumard, 2008).

Nous appellons, par facilité, cette première forme de raisonnement artificiel « l'intelligence de production ». Elle a pour racine les deux premières révolutions

industrielles, le management scientifique et la cybernétique pour sa forme contemporaine. Mais on la retrouve dans toute formulation mathématique dont l'objectif est de permettre à un ensemble humain de pouvoir accomplir des tâches répétitives tout en améliorant continuellement leur performance. Le code d'Hammourabi est sans doute l'une des premières expressions de cette intelligence de production. Il permet par de simples équations à des êtres humains de systématiser les récoltes, de régler des différends sur du bétail, de calculer le partage de l'impôt. Il s'agit bien d'un code, d'un ensemble d'instructions, gravé dans la pierre, permettant une intelligence de production répétée et continuellement améliorée (par l'adjonction de nouvelles instructions ou règles dans le code d'Hammourabi).

Une règle est en soi une forme d'intelligence désincarnée. Toute forme arithmétique généralisée, tout système d'équation, depuis et avant l'invention de l'algèbre par Khwârismî au début du IX$^{e}$ siècle, sont des expressions d'une intelligence désincarnée, autonome, qui peut être répliquée et reprise à son compte par n'importe quel être humain. Newell (Newell et Simon, 1972, 1976 ; Newell, 1983) ne dément pas cette perspective en affirmant que l'histoire de l'intelligence artificielle débute au moment où l'on commence à séparer des fonctions de représentation, de résolution, de reconnaissance et d'acquisition de connaissances ; ce qui l'amène à considérer la naissance de l'intelligence artificielle, en tant que discipline, dans l'opposition entre l'idée de mécanisme artificiel et de téléologie dans la période 1640-1945. Il s'agit bien sûr de l'opposition cartésienne entre l'esprit et la matière, qui pour Newell aboutit aux interrogations de la cybernétique à la fin des années 1940. L'idée de la cybernétique est de proposer qu'une machine capable d'une réponse autonome, un *feedback*, était *de facto* incarnée d'un propos (Rosenbleuth *et al.*, 1943 ; McCulloch et Pitts, 1943).

Une des idées fondatrices de la cybernétique était donc que l'intelligence présuppose un propos ; ce qui pouvait simplement se tester par la capacité d'une machine à « accomplir des tâches d'une difficulté grandissante ». Deux problèmes se posaient alors : chez l'être humain, le *feedback* intègre la problématisation aussi bien que la réponse ; le *propos* impliquait la notion anthropocentrique de conscience, dont les machines sont bien sûr dépourvues en 1950.

Miller *et al.* (1960, p. 26) proposèrent de substituer la notion *d'incongruité* à celle de *feed-back*. Dans leur célèbre exemple d'apprentissage expérimental, c'est l'incongruité d'une réponse reçue qui motive l'homme à s'acharner à trouver une

réponse : le « propos » est donc bien encapsulé dans la forme du *feedback* (Jones, 1975). Ils en déduisent que la réaction d'un effecteur à un événement dépend de la réalisation d'essais, puis d'un nouveau test des procédures opérationnelles tout en observant comment ces itérations « modifient le résultat de l'essai » (p. 25) : « L'action est initiée par une "incongruité" entre l'état de l'organisme et l'état testé, et l'action persiste jusqu'à ce que l'incongruité (c'est-à-dire le *stimulus proximal*) soit éliminée. » (*op. cit.*, p. 26).

Ils ont appelé ce modèle initial « l'hypothèse cybernétique », mieux connu aujourd'hui sous le nom du principe de « TOTE » (*Test-Operate-Test-Exit*). L'approche TOTE est un algorithme commun pour résoudre des problèmes non déterministes au sein d'une conception complexe, suivant les étapes de test afin d'obtenir une représentation du problème, opérant pour créer un changement qui peut être caractérisé, testant à nouveau afin d'observer le comportement du résultat, et sortant lorsque le problème est résolu. Miller *et al.* s'intéressaient à « l'utilisation de l'ordinateur comme automate pour illustrer le fonctionnement de diverses théories psychologiques » (*op. cit*, p. 50).
17Comme le soulignait Newell (1982), posséder un propos artificiel ne résout pas le problème de l'existence d'une intelligence artificielle. Cette perception anthropocentrique de l'intelligence artificielle comme une abstraction désincarnée de fonctions cognitives (l'apprentissage et la résolution de problèmes) est dominante pendant la seconde moitié du XIX$^{e}$ siècle (dans la science-fiction et la littérature fantastique) jusqu'aux premières tentatives de modélisation de la psychologie humaine au début du XX$^{e}$ siècle.

*Les lois de la pensée* de Boole (1854) sont sans doute l'œuvre la plus sous-estimée pour l'anticipation de l'intelligence artificielle contemporaine ; c'est-à-dire celle dominée par le paradigme connexionniste du début du XXI$^{e}$ siècle. Boole écrit : « Les lois générales de la Nature ne sont pas, pour la plupart, des objets immédiats de perception. Elles sont soit des inférences inductives à partir d'un large corps de faits, la vérité commune dans laquelle elles s'expriment ; soit, dans leur origine au moins, des hypothèses physiques d'une nature causale permettant d'expliquer des phénomènes avec une précision sans faille, nous permettant ainsi de prédire de nouvelles combinaisons de celle-ci. » (Boole, 1954, p. 4). Boole exprime dans les lois de la pensée les futurs fondements de l'apprentissage profond. Il considère qu'il n'existe qu'une seule loi naturelle, celle de « conclusions *probables* » (l'italique est de l'auteur) qui s'approchent de la certitude au fur et à mesure où elles reçoivent une confirmation par l'expérience.

L'intuition booléenne s'étend, dès 1854, dans la conviction que l'ensemble des problèmes humains puisse être résolu par un traitement logique et symbolique. Boole, bien sûr, est un enfant de son siècle, fortement influencé par le rayonnement des théorèmes de Thomas Bayes, de Richard Price ou de Pierre-Simon de la Laplace, qui dès la fin du XVIII[e] siècle avaient construit la théorie des probabilités. Mais la différence est que Boole offre une réflexion synoptique, où il pose, dès le II[e] chapitre, le problème du langage et de la représentation (p. 25-28) où il définit les signes comme des « marques arbitraires » ; pour s'intéresser ensuite à un découpage systématique des fonctions cognitives humaines (« les principes du raisonnement symbolique »). Bien que son vingt-deuxième chapitre soit une reddition (« C'est une capacité inhérente de notre nature que d'apprécier l'Ordre (..) » (*op. cit*, p. 403), ou peut-être simplement un acte de prudence, Boole, le premier, fait de la logique un système d'investigation du réel qui préfigure l'informatique heuristique des années 1950 (notamment avec ses principes d'interprétation, d'élimination et de réduction). Newell (1982, p. 8) identifie dès lors une des caractéristiques dysfonctionnelles de l'intelligence artificielle : « La logique est devenue la technologie sous-jacente qui permettait à des mécanismes d'accomplir des choses. De fait, c'est précisément la séparation entre la logique et la pensée qui a permis à la logique de devenir une science de jetons insignifiants manipulés selon des règles formelles, qui à son tour a généré la pleine mécanisation de la logique. » Newell identifie la fin de cette séparation entre pensée et logique quand McCulloch et Pitts (1943), s'inspirant des travaux de Turing, suggèrent que les liaisons synaptiques du cerveau pourraient fonctionner à l'image d'un réseau de probabilités causales selon un anneau de Boole (cf. figure 1).

**Figure 1**

**Représentation des connexions nerveuses (*nerve-net*) par McCulloch et Pitts (1943, p. 130)**


Deux perspectives s'affrontent déjà dans ces préfigurations : d'une part, celle d'une « mécanisation de la pensée », c'est-à-dire la capacité à reproduire un

raisonnement au sein d'une machine ; et d'autre part, celle d'une intelligence du dépassement des limites inhérentes à la pensée humaine.

Cette seconde perspective va être accélérée par les défis posés par le Seconde Guerre mondiale. Le premier défi est celui de casser le chiffrement allemand, défi relevé par Alan Turing avec succès ; mais c'est sa publication de 1936 sur la notion de « calculabilité » (inspirée d'A. Church), pour laquelle Turing utilisa le terme de « computability » qui sera décisive ; c'est-à-dire le problème des nombres dont les décimales sont calculables par des moyens finis. Loin d'être une réflexion philosophique, l'article de Turing était le descriptif d'une machine (un algorithme, ce que sont *in fine* les machines de Turing) dont le mouvement était partiellement déterminé par une configuration initiale (l'ensemble des possibles). Ainsi l'écrit Turing : « Si une *a*-machine imprime deux sortes de symboles, dont le premier (appelé formes) consiste entièrement de 0 et de 1 (les autres étant appelés symboles de second ordre), alors cette machine peut être appelée une machine de calcul » (Turing, 1936, p. 232) (*computing machine* dans le texte original ; le « mouvement » renvoyant figurativement à l'idée du cylindre de Leibniz de 1673). Turing essaye bien sûr de répondre au challenge de Hilbert et Ackermann de 1928, connu sous le nom du « Entscheidungsproblem » (le « problème de la décision » en allemand) qui consiste à déterminer de manière algorithmique s'il peut être dérivé par un système de déduction sans autres axiomes que ceux de l'égalité. Afin de démontrer l'indécidabilité arithmétique de la proposition de Hilbert et Ackermann, Turing crée une machine (un algorithme) qui prend en entrée un énoncé d'une logique du premier ordre et répond « Oui » ou « Non » selon que l'énoncé est universellement valide, c'est-à-dire valide dans chaque structure satisfaisant les axiomes.

La machine de Turing (qu'il appelle l'a-machine, pour « machine automatique ») inverse astucieusement la question posée. On ne cherche pas directement à prouver (par la négative) Hilbert et Ackermann, mais on génère, de manière continue, autant de machines possibles jusqu'à ce que la preuve soit apportée ; c'est-à-dire des machines capables de produire des calculs arbitraires. Ce n'est pas, bien sûr, réellement une machine, et le « calculateur », en 1936, est Alan Turing lui-même ; mais Turing ouvre la voie à la naissance de l'informatique de l'après-guerre, même si un ordinateur n'est fondamentalement pas capable de créer une infinité de machines, l'article de 1936 contient à la fois la conception initiale d'une architecture de calculateur, ainsi que la base logique d'un langage de programmation.

Quand les premiers ordinateurs apparurent dans les années 1940, ils étaient lourds, lents et à la précision fortement douteuse. Les premières machines, héritières de Turing, étaient analogues ; ce qui leur donnaient une vitesse bien supérieure aux premières machines numériques. Ce changement fut si rapide, que dès le début des années 1950, le monde scientifique devient littéralement fasciné par l'exploration des possibilités qu'offraient de telles puissances de calcul (nous parlons bien sûr ici de « puissances » qu'une simple calculatrice peut aujourd'hui égaler et dépasser… par un facteur 1000). Jusqu'alors, ce qu'on appelait des « ordinateurs » étaient de larges feuilles de calcul automatisées dont la seule fonction était de résoudre des équations différentielles partielles. Il n'y avait pas de demande spécifique dans la société pour une autre forme de calcul. Après tout, le code Hammourabi avait bien géré la société mésopotamienne avec des règles de trois, et l'Empire Byzantin s'était bien répandu sur la terre avec une algèbre assez sommaire. Aucune des formes de calcul consommée par la société d'après-guerre ne peut recevoir le qualificatif « d'intelligence artificielle », et encore moins ce qui est fait à l'époque dans les laboratoires universitaires (Simon, 1991).

Quand William H. Starbuck (1981) défie le groupe de chercheurs réunis autour de Pugh dans l'école (naissante) de la théorie de la contingence, la première chose qu'il demande est qu'on lui donne accès aux quarante cartons de données qui sont des cartes perforées. Les « ordinateurs » servent l'effort scientifique en réalisant des calculs qui auraient été fastidieux, voire incommensurables, s'ils étaient réalisés à la main. Comme l'écrit Simon : « Bien sûr, ce n'est pas de l'heuristique parce qu'il y a un théorème qui prouve que vous arriverez tôt ou tard au résultat, et peut-être est-ce pour cela que l'on ne peut pas le considérer comme de l'intelligence artificielle. » (Simon, 1991, p. 128). Quelque part, la machine de Turing est capable de se multiplier pour résoudre un problème, ce que la logique dominante de l'informatique, créée pour faire du calcul servile et automatiser des tâches, est à mille lieues conceptuelles.

Ainsi, les premières expressions de l'intelligence artificielle sont essentiellement performatives. La philosophie du XIX$^{e}$ siècle établit le caractère associatif de la pensée humaine ; un des premiers langages de programmation logique, le LISP, repose sur une structure associative. Si nous avions eu une philosophie occidentale dominée par l'idée d'imprégnation organique, et non pas d'association cartésienne, les tentatives allégoriques des années 1950 auraient pu être intéressantes ! L'abstraction des premières machines capables de « résoudre des problèmes » ne s'éloigne jamais de ces allégories, inspirées de la biologie, de la psychologie ou de la philosophie. Ces premières abstractions, comme le note Simon

(*ibid.*) sont plus rhétoriques qu'heuristiques. La structure des langages de programmation comme LISP, mais également les premières expérimentations comme le perceptron (Rosenblatt, 1956, 1957a, 1957b ; Minsky et Papert, 1969), reposent toujours sur le même principe que les êtres humains partent de problèmes et vont vers des solutions. C'est comme si le trait de génie de Turing (générer autant de machines que possible jusqu'à ce qu'une trouve capable de générer le problème qui satisfasse la solution) ou l'intuition fabuleuse de Boole (le monde est fait de conclusions probables) soient passés à l'as par ce que la population mondiale comptait de génies au XX$^{e}$ siècle.

La structure dominante (de la conception de machines à résoudre des problèmes) devint rapidement la séparation systématique entre « tout ce que l'on sait » (propositions, base de données, entrées) et tous les opérateurs logiques pouvant s'y appliquer. En d'autres termes, un paradigme de l'apprentissage « stationnaire » s'installe dès la conception initiale des premiers apprentissages artificiels.

Essayons d'en détailler les principes. Trois impulsions décisives façonnent, autour d'Herbert Simon, la naissance de l'intelligence artificielle après-guerre (1945-1958). La première répond à une fascination de la montée en échelle de la puissance de calcul, et l'objectif est de construire des programmes informatiques capables de démontrer *n'importe quelle* forme d'intelligence. Ce mouvement est assez similaire à l'engouement des années 1990-2020 pour les réseaux de neurones et l'apprentissage profond. Il n'y a aucune révolution architecturale ou paradigmatique ; mais on peut soudainement repousser les limites physiques du brassage de données. En d'autres termes, la révolution de l'apprentissage profond (*deep learning*) est celle d'enfants à qui l'on vient de confier des jouets à la puissance démesurée (littéralement) ; mais qui, à l'instar des pionniers de l'intelligence artificielle, n'ont pas de *propos* particulier pour leur usage. Dans la même logique que celle de Turing (1936), ces apprentissages profonds permettent d'entrer dans des logiques abductives, au sens de Blaug (1982), où il n'existe qu'une limite énergétique et financière à la recherche de combinaisons du réel pouvant expliquer les combinaisons probabilistes qui émergent des apprentissages successifs des couches neuronales. L'apprentissage profond n'est en somme qu'une formidable computation capable de reconnaître les éléments d'un réel *préexistant* ; c'est-à-dire qu'il peut, grâce aux capacités de calcul contemporaine, trouver ces « conclusions probables » (Boole, 1854), ces modèles combinatoires sous-jacents qui expliquent le mieux, probablement, les éléments épars et saillants isolés par les filtres successifs des apprentissages bruts.

La deuxième « poussée » de l'après-guerre est celle initiée par la Rand et le Carnegie Institute of Technology, dont le nom original était Groupe d'étude du « complex information processing » (Simon, 1995, p. 96). L'objectif était, sur le plan national américain, de produire avant le reste du monde une théorie des systèmes intelligents. Un des effets collatéraux de ce programme de la Rand fut la conduite des travaux menés sur la décision humaine par Jim March et Herbert Simon, qui aboutirent à la publication *Des organisations* (1958), et créèrent les pierres fondatrices des sciences administratives (que l'on appelle aujourd'- hui sciences de gestion).

La troisième impulsion, toujours dans la périphérie des financements nationaux de la RAND, de la Défense américaine et du programme présidentiel de reconstruction (où Herbert Simon siégeait) était de pouvoir « désincarner » cette intelligence, c'est-à-dire d'en faire des systèmes-experts pouvant accomplir des tâches humaines, ou conduire des véhicules, assurer un système de défense, protéger la stabilité d'un système de décision.

Pour Simon (1995, p. 97), ces trois agendas (comprendre l'intelligence, comprendre la cognition humaine, extraire dans un système l'intelligence humaine) créèrent trois communautés scientifiques très distinctes. La première se consacrait au développement d'une théorie de l'intelligence ; la seconde inventa les sciences cognitives (une vision fonctionnelle et mécaniste de la perception humaine, qui réunissait psychologues, anthropologues et informaticiens) ; et une troisième se consacra à explorer les capacités de vérification et de preuves théoriques que pouvait apporter l'IA aux mathématiques, à la cryptographie, à la science du code.

## II – L'enfermement stationnaire ou quand les machines échouent à désapprendre

Mais une chose fut commune à ces trois groupes sociaux : leur usage des ordinateurs. Ce que Turing en 1936 conçoit comme une « machine » abstraite est désormais rendu possible par la programmation. La recherche devient fortement consommatrice d'échantillons aléatoires, de simulations et séries stochastiques. Lorsque l'on veut étudier un phénomène, il est utile de le considérer comme le produit d'un processus ou d'une distribution aléatoire ; et il est critique de pouvoir « isoler » le modèle, lorsqu'il émerge, ou lorsqu'il est postulé, en rendant ses espérances mathématiques indépendantes du temps.

Par exemple, si je souhaite étudier la marée bretonne, et que je souhaite modéliser les vagues de la Côte d'Armor, qui sont évidemment les plus belles de France, je peux supposer qu'il n'existe pas une infinité d'expressions d'une vague ; c'est-à-dire qu'au bout d'un moment d'observation, je pourrai très bien anticiper la forme de la prochaine vague, ou l'apparition d'un ressac. On assume ici que la « moyenne temporelle » est à peu près identique à la moyenne d'ensemble. On parle alors d'un *processus stationnaire* ; l'adjectif exprimant ici que, dans son ensemble, le phénomène est considéré « stationnaire », et se reproduira à l'identique. En d'autres termes, on isole le modèle de la variable temps (*akheros* en grec moderne ; littéralement, « en dehors du temps »).

D'un point de vue purement technique, un ordinateur est une machine stationnaire, telle qu'imaginée par ses concepteurs, et nous parlons ici autant d'un point de vue philosophique qu'ingénierique, soit Leibniz, Babbage, Lovelace, Boole, etc. La structure sous-jacente du modèle n'évolue pas avec le temps. Un ordinateur ne change pas de modèles quand il reçoit un signe d'échec d'une opération. Il informe l'utilisateur que l'opération a échoué. Et c'est fort heureux : il serait bien difficile de gérer un ordinateur dont les propriétés changeraient dès que l'on change son « repère temporel ». Le comportement d'un système d'information doit être déterminé, et globalement, déterministe. « I am sorry, Dave, I'm afraid I can't do that », la réponse perverse de HAL dans *A Space Odyssey* de Kubrick (1968), ne fait pas partie, en dehors des films de science-fiction, de la population des comportements attendus d'un système.

On retrouve différents degrés de stationnarité dans de nombreux développements des probabilités et de l'intelligence artificielle. La stationnarité « absolue » serait l'observation d'un comportement toujours identique quel que soit le point temporel $t$ ou $t + n$ auquel est faite cette observation (Widrow *et al.*, 1976)[2].

Le qualificatif de « stationnaire » signifie qu'il existe une grande probabilité de retrouver un état donné qui soit une constante au cours de temps (la forme de la vague bretonne), même s'il existe de nombreuses transitions entre les différents états, plus ou moins imprévisibles (« des sauts d'état »). Un état stationnaire est donc une loi d'équilibre dynamique, postulée, promulguée (par échantillonnage) ou observée.

[2] C'est une notion importante lorsque l'on traite de séries temporelles. Si l'on utilise des séries temporelles *non stationnaires*, on peut générer des régressions linéaires dont les résultats sont artificiels et erronés. On parle alors de « régression fallacieuse » (« spurious regression », cf. Granger et Newbold, 1974).

La plupart des activités de recherche se conduisent par la médiation de modèles stationnaires ; notamment parce que nous recherchons la régularité d'un phénomène, pour exprimer son caractère reproductible. Lorsqu'un chercheur mène une recherche qualitative, il va émettre, depuis l'état de l'art, ou les faisant émerger, *tabula rasa*, à partir d'une observation ethnographique, des « catégories d'observation ». L'idée même de « catégories d'observation » dont on recherche la régularité appartient à une conception stationnaire de la modélisation. Il ne s'agit pas d'un principe d'invariance. Il s'agit d'exprimer que, quelle que soit la forme particulière *n* que peut prendre l'observation de ma catégorie (par exemple, l'extension du périmètre d'action d'un dirigeant), l'existence de cette catégorie ne dépend pas de *n*.

Le raisonnement est tout aussi valide pour une méthodologie quantitative. Nous modélisons à partir d'abrégés et de réductions, qui sont soit les théories existantes, soit des relations que nous découvrons par une étude qualitative, ou éventuellement, en procédant à différentes triangulations (Jick, 1979).

Un apprentissage « stationnaire » est donc un apprentissage qui, par facilité ou commodité de calcul, va « figer » un échantillon pertinent (former une tendance) ; c'est-à-dire se concentrer sur l'extrait du phénomène dont le motif semble se répéter dans le temps. Par extension, on peut considérer que tout apprentissage qui, à partir d'un modèle antécédent, ou d'un modèle « cible » *ex post* (en pensant à l'apprentissage profond), apprend l'évolution des variables attachées à ce modèle, sans questionner la permanence ou l'invariance du modèle sous-jacent, est un apprentissage stationnaire[3] phénomène non stationnaire, dans une série temporelle, n'affichera pas de tendance reconnaissable (moyenne et variance ne sont pas constantes dans le temps). Un apprentissage non stationnaire ne cherchera pas à maintenir une tendance, un schéma ou un modèle sous-jacent invariant du temps. Par exemple, lorsque l'on veut détecter une anomalie comportementale dans un système, on peut supposer que « n'importe quel programme ou application peut accepter un large éventail d'entrées, communiquer avec d'autres programmes, et les chemins d'exécution choisis (où le processus est localisé et ce qu'il y fait) sont souvent au moins dépendant des entrées » (Hourbracq *et al.*, 2018, p. 3). Il n'y a pas de comportement « moyen » *a priori*, et le but n'est donc pas de tester ou de prouver des modèles émis à l'avance, mais de découvrir ces modèles de façon automatique. Avec

[3] Nous incluons donc ici des familles d'apprentissage qui ne répondent pas à la définition mathématique stricte de la stationnarité, c'est-à-dire des processus stochastiques dont la distribution inconditionnelle des probabilités communes ne change pas dans le temps.

un certain abus de langage, on peut estimer qu'un apprentissage non stationnaire découvre et apprend des modèles, pas des variables.

La cognition humaine est profondément non stationnaire. Starbuck (1963) douta très rapidement des modèles qui supposaient une séquence figée entre la découverte d'un problème et la recherche, plus ou moins optimale d'une solution. Dès 1963, il questionne à la fois la notion de « satisficing » de Simon, les mesures de Siegel et la théorie des aspirations (attentes) de Festinger. Ce que dit Starbuck est très simple : les êtres humains ne maximisent pas leurs attentes d'une manière rationnelle (Festinger), pas plus qu'ils ne se contentent systématiquement d'un niveau satisfaisant (Simon). Dans le choix de nos alternatives, c'est-à-dire essayer d'atteindre nos buts au maximum, réduire nos ambitions, ou inventer un but de substitution, nous procédons par inférences ; nous ne percevons pas toutes les conséquences de nos choix ; cette perception n'est pas séquentielle ; nos définitions subjectives du succès deviennent de plus en plus malléables au fur et à mesure que ce succès s'éloigne. Notre imagination est capable de générer des modèles superflus, placebos, réellement distants de la réalité observée, mais qui nous permettent de maintenir notre flux d'action, ou notre ancrage dans la société (si on choisit une perspective plus institutionnelle). Bref, nous changeons la « vague » selon les circonstances, nous « gelons » une idéologie parce qu'elle nous permet de poursuivre une théorie déjà démise par les faits. Nous construisons des modèles qui peuvent expliquer nos données ; parce qu'il y a peu que nous puissions faire à propos de ces données.

Du point de vue l'intelligence artificielle, un tel apprentissage serait non stationnaire, autonome et non supervisé[4]. Nous passons nos journées à théoriser en générant des explications et des modèles très approximatifs de ce que nous observons. En d'autres termes, nous sommes des machines à produire des lois normales fantasques à propos de n'importe quoi, et n'importe quand. Dire que la rationalité humaine est « limitée », à cet égard, fait preuve d'un solide optimisme.

Il est intéressant de remarquer, à ce stade, que nos actes de théorisation scientifique ont peu en commun avec le fonctionnement commun de notre cognition. Si ces « approximations fainéantes » ne sont pas à proprement parler des théories robustes (Weick, 1995), elles sont néanmoins l'arc sous-jacent de toute

[4] Le paradoxe humain est que nos comportements, de leur côté, sont profondément « stationnaires », au sens où la programmation comportementale que nous nous imposons en tant qu'êtres humains est un élément vital de notre survie individuelle et collective. Les routines, les saisons, les rituels, les institutions, les *habitus*, offrent un réceptacle comportemental stable au foisonnement erratique de notre cognition.

théorisation : un diagramme, un schéma griffonné sur un coin de table, sont autant de quasi-théories, que nous éprouvons ensuite par la méthode scientifique. Une chose que, en tant qu'êtres humains, nous sommes incapables de faire est d'apprendre des modèles sans les connaître. Nous pouvons générer des combinaisons de modèles qui peuvent expliquer un phénomène ; c'est-à-dire effectuer des sauts inductifs informels, imaginatifs, créatifs, entre des preuves éparses et une théorie explicative inventée à la volée (la fameuse « abduction » de Blaug, 1982) ; Mais contrairement à une machine (ici au sens d'algorithme), nous ne pouvons être complètement « non stationnaires » : nous recherchons toujours à imposer une micro-explication stationnaire, une loi normale temporaire (une intuition), puis une seconde, puis la combinaison de plusieurs de ces lois temporaires. Nous sommes tellement « durcis » par ces microthéories professées que le seul moyen de réellement s'en séparer est de désapprendre ; en se débarrassant d'abord des mythes rationnels, pour finalement se débarrasser du « modèle » (Hedberg, 1981).

La plupart des apprentissages machine ne sont pas très différents des apprentissages humains, parce qu'ils en sont tout simplement inspirés. L'inspiration derrière l'invention de l'apprentissage profond repose sur l'idée que le cerveau animal ou humain dispose d'architectures profondes, constituées de réseaux de synapses qui réalisent le travail d'essai-erreur-apprentissage de production de « théories qui fonctionnent ». L'idée fut dès lors, notamment sous l'impulsion de Yann LeCun et de Françoise Fogelman-Soulié (1987), mais grâce aux travaux d'Alexey Ivakhnenko (1982) sur l'apprentissage statistique inductif ; de recréer cette capacité cognitive humaine en reproduisant ces différentes couches de filtres synaptiques, capables de reconnaître des caractéristiques éparses d'une image, pour la deviner ou la « désigner » par induction formelle.

## III – L'acte de découverte scientifique : de l'humain vers la machine

Peut-on, pour autant, parler de « machines intelligentes » ? Chez Ivakhnenko (1968), il y avait le désir de créer une nouvelle philosophie de la recherche scientifique. Alexei Ivakhnenko cherche une solution à la modélisation mathématique de processus dont on n'a aucune connaissance *a priori*. Il propose une approche, qu'il appelle l'appréhension des données par le groupement (*Group Method of Data Handling* ou GMDH), qui consiste à assumer que le résultat comportemental $y$ d'un système inconnu, se comportera comme la fonction de ses $m$ valeurs d'entrée. La proposition théorique d'Ivakhnenko était elle-même

inspirée des premières recherches sur les perceptrons et le principe d'auto-organisation (Madala et Ivakhnenko, 1994).

L'objectif était de formaliser une méthode inductive permettant de résoudre des problèmes de reconnaissance, de modélisation et de prédictions de processus aléatoires. Mais l'idée d'Ivakhnenko était surtout de créer une nouvelle « science » de la résolution de problèmes complexes, « unifiant les théories de la reconnaissance de formes et le contrôle automatique dans une nouvelle métascience » (*ibid.*, p. III). L'idée du GMDH était de construire des modèles mathématiques offrant une approximation des formes (*patterns*) inconnues de l'objet ou du processus étudié. Les composantes de cette approche sont la génération automatique de modèles, l'utilisation dans la modélisation de décisions non conclusives (à l'instar du fonctionnement réel d'une cognition humaine), accompagné d'une sélection par paires des critères (prémices) permettant une explication externe finale optimale du système complexe inconnu. Cette procédure est celle qui est utilisée aujourd'hui, en 2020, dans les modèles de *deep learning* (apprentissage profond). Utilisant son approche GMDH, Ivakhnenko construisit le premier réseau de neurones profond à huit couches en 1971 ; faisant de lui le créateur du premier apprentissage profond.

L'opportunité qu'Ivakhnenko essayait de saisir était la formidable capacité des machines (calculateurs, ordinateurs) à systématiser l'exploration de modélisations par l'extraction directe de connaissance à partir d'un dispositif expérimental ; ce qu'il appela tour à tour des « mécanismes de prédiction cybernétique » ou encore des modèles d'auto-organisation. À la différence d'un être humain, une machine peut « épuiser » les tests de possibilités de relations (causales, de covariance, de mutualité d'information, etc.). N'y avait-il pas là un moyen de faire de la « science » différemment ? (Mitroff et Kilman, 1977) Nous englobons communément dans le terme de « science » tout autant l'idée de corpus de connaissances, l'activité humaine produite par des scientifiques que l'institution sociale émettant et contrôlant sa légitimité et son statut. Ce faisant, nous créons un amalgame entre l'activité de découverte scientifique et son mécanisme sociétal de validation (Ziman, 1968, p. 11). Cette distinction entre le contexte de la découverte et le contexte de justification, introduite par Reichenbach (1938), prend une dimension nouvelle dans la perspective d'une intelligence artificielle impliquée dans une production scientifique. Pour Reichenbach, le « contexte de la découverte » est l'ensemble des activités de recherche conduites en milieu naturel, opposant ainsi l'expérience à la prédiction ; la connaissance naturelle, empirique et positive, à son test et à sa mise à l'épreuve (le « contexte de justification »).

Le bon sens de ce postulat ne tarda pas à faire de Reichenbach un auteur légitime (Schiemann, 2006, p. 24), et d'installer durablement l'idée d'une opposition, dans l'acte scientifique, entre une phase empirique de recueil de données, et une phase logique de tests et de validation. C'est ce que nous appellons « la part humaine » dans la découverte scientifique : notre questionnement, nos problématisations, qui peuvent ressembler à une génération automatique de modèles approximatifs (Weick, 1995 ; Ivakhnenko, 1968) sont toujours les dérivés d'une implication sociale : dans une communauté scientifique, qui nous juge, nous paye, nous promeut ou nous dégrade ; dans une société, préoccupée par des questions triviales, comme par exemple, le remplacement « machinehomme ».

Une critique évidente de la proposition de Reichenbach est qu'il est difficile de séparer ce qui appartient au contexte et ce qui appartient à une modélisation abstraite. Notre recueil de données, notre appartenance au monde naturel, n'est pas un ensemble mécanique. Si bien que la « reconstruction rationnelle », qu'il emprunte à Rudolf Carnap (1928), ne peut réellement atteindre cet « ordre optimal » où les observations et leurs conséquences sont rangées dans un système cohérent (Reichenbach, 1938, p. 5). En imposant une dichotomie entre un acte « neutre » de recueil, fondé uniquement sur l'ordonnancement logique de ses étapes, et un activité d'abstraction lui étant ultérieure, Reichenbach installe un paradigme qui deviendra dominant dans les sciences contemporaines : celui de la modélisation stationnaire et du cycle « modélisation – tests – approbation », qui précepte l'idée de modèles à celui de données.

Un réseau de neurones est familier de la conception séparatiste présentée par Reichenbach. Le « contexte de la découverte » y est un processus de recueil ordonné et logique. Une capture de données recueille les plus infimes parties du réel observé, et essaye ensuite, dans un « contexte de justification », et par induction formelle, de rapprocher de manière probabiliste chacun de ces éléments à une forme déterministe appartenant à un autre ensemble de probabilités conduisant, par tests successifs, à la reconnaissance probabiliste d'une forme ou d'une réponse vraisemblable. Le réseau de neurones est respectueux du dogme fondateur de l'empirisme logique : les données brutes ne font pas l'objet d'une interprétation singulière. L'outlier (l'élément discordant dans le nuage de points) ne fait pas l'objet d'une exploration *ex ante*.

L'ancêtre des algorithmes d'apprentissage profond contemporains, le perceptron de Frank Rosenblatt (1957) fut le premier à proposer une substitution

machine à l'apprentissage humain. Il était fondé sur un principe de séparabilité statistique, permettant, d'après Rosenblatt, « la reconnaissance de patterns complexes avec une efficacité bien plus grande que les ordinateurs courants ». Ceci est écrit le 3 avril 1957 : « Les dispositifs de ce type sont appelés à être capable de formalisation, de traduction de langage, de recueil de renseignement militaire et capables de résoudre des problèmes par l'induction logique » (Rosenblatt, 1957b, p. 1).

Mais dès les années 1980, le propos de l'apprentissage machine se sépare en deux voies très distinctes. La voie de « l'heuristique augmentée » ou de la découverte automatique de modèles (Turing à l'origine en 1936, puis Ivakhnenko en 1965) abandonne clairement la partie au profit de l'école de la reconnaissance (audio, vidéo, images). Deux éléments peuvent peut-être expliquer cette scission. Le premier est la disponibilité de la donnée. L'extraction d'images, telle qu'elle est préfigurée par Fukushima et son « neocognitron » (1980) permet une production infinie de dispositifs expérimentaux ; la donnée pouvant être un enregistrement audio, une banque d'images. Le second tient à la transformation de nos sociétés contemporaines et des pratiques de la recherche scientifique.

La disponibilité de la donnée est accompagnée d'une forte demande des financeurs, notamment les agences de recherche de la défense américaine, qui entrent dans l'ère d'une écologie martiale numérique où la dominance stratégique s'affiche déjà, dès 1980, comme une vectorielle de la maîtrise de l'interprétation de données (satellitaires, militaires, communication). Toutes les applications de reconnaissance, poussées par ces financements, s'épanouissent fortement dans la décennie 1980-1990, que ce soit dans la reconnaissance de formes, la synthèse vocale ou la prédiction verbale. C'est d'ailleurs un financement du Bureau pour la recherche navale (Office of Naval Research – ONR) qui permit à Rumelhart *et al.* (1986) de proposer leur modèle de rétro-propagation (« back-propagation ») qui montre qu'un réseau de neurones peut apprendre à filtrer son apprentissage.

Les premières applications sociétales advinrent rapidement : les réseaux convolutionistes de Yann LeCun (1987) permettaient la reconnaissance de l'écriture manuscrite de près de 20 % du volume des chèques traités aux États-Unis au début des années 1990 ; jusqu'à ce que l'on réussisse, notamment avec les travaux de Schmidhuber, de Cortes ou de Vapnik à étendre à un très grand nombre (1 000 +) la possibilité de couches successives d'apprentissage. Ce qui était une communauté scientifique devint très rapidement un paradigme industriel ; avec la naissance du

cyberespace à la fin des années 1990 ; et la transformation numérique de la plupart des économies occidentales (1998-2020) qui permet de monétiser l'apprentissage profond en réalisant une segmentation marketing *inverse* (rapprocher un client d'un segment préexistant que l'on vise). Google, Facebook, Amazon firent de ces sauts inductifs par *deep learning* le levier de leur croissance mondiale.

La recherche en « intelligence artificielle » devint obsessionnelle autour de trois concepts : les modules d'entraînement des réseaux de neurones (« Trainable classifier modules ») ; les modules d'extraction de données ; et l'accumulation de données exploitables. Un régime stratégique de la donnée (« Data drives learning ») s'installa brutalement et durablement, en préparant des sets de données gigantesques pré-indexés permettant d'accélérer l'entraînement, mais surtout de vérifier sa pertinence (l'ImageNet de Fei-Fei Li à Stanford avec ses 14 millions d'images indexées) ; jusqu'à l'expérimentation du « chat » en 2012, de J. Dean et A. Ng (Stanford/Google) consistant à être capable de reconnaître de façon non supervisée un chat à 16 000 processeurs et 20 millions de vidéos YouTube.

## IV – L'induction est-elle réellement mon amie ?

Mark Blaug (1982, p. 12-25) offrait en 1982 une réflexion pertinente sur le « problème de l'induction », qui peut se résumer ainsi : nous sommes limités par le nombre d'observations que nous pouvons humainement conduire ; ce nombre est si restreint que cette poignée de faits bruts constitue déjà des théories ; et ainsi, « il n'y a pas de logique de la découverte et il n'y a pas d'avantage de *logique* démonstrative de la justification » (p. 25). Blaug proposa dès lors une forme de positivisme aménagé (Koenig, 1993) consistant à accepter l'idée qu'il existe une réalité objective testable ; et que celle-ci pouvait être atteinte par abduction, c'est-à-dire par ces fameux sauts inductifs réalisés, aujourd'- hui, par les réseaux de neurones d'un apprentissage profond, qu'il soit en rétro-propagation ou en apprentissage adverse (apprenant contre lui-même).

L'apprentissage profond est supposé résoudre les limites inhérentes à la limitation cognitive individuelle ; incluant le risque de « pré-théoriser » puisqu'un réseau de neurones cherche aveuglément toute combinaison possible. Le premier écueil est que le sous-espace vectoriel de la pensée est très limité par rapport à la pensée elle-même. Quelle soit prise dans un sens hypothéticodéductif ou dans le sens inverse (inductif ou abductif), la pensée humaine ne peut être réduite à un problème

connexionniste. Si l'apprentissage profond devait aboutir à une reproduction du sens commun, il aboutirait à une émulation parfaite de l'esprit grégaire.

On pourrait imaginer que l'intelligence artificielle pourrait intervenir en amont de ces sauts imprévisibles entre intuition et découverte ; c'est-à-dire en aidant le chercheur à trouver les combinaisons les plus contre-intuitives (au sens de Davis, 1971) ; c'est-à-dire à lutter justement contre le biais institutionnel qui fait que les chercheurs citent plus souvent les chercheurs réputés et notoires, en comprenant très rarement le sens de la paternité scientifique, et en essayant de maximiser le ratio d'impact de leurs citations. Mais une telle hypothèse ne serait efficiente que si nous séparons l'égo de la création scientifique ; c'est-à-dire « désindividualiser » les processus de découverte où l'égo du scientifique (ou la bataille d'égos) peut devenir toxique, comme une revue de la littérature (qui est avant tout une sélection politique et un calcul de positionnement), l'établissement d'un argumentaire pour ou contre un état de l'art ; les choix « opportuns » d'oublier la paternité d'un concept, en le reprenant à son compte pour ensuite invoquer qu'une pure coïncidence ait généré un plagiat involontaire.

Le seul moyen serait de créer des « magmas connexionnistes » permanents, vivants, non supervisés et non stationnaires : des états de l'art en devenir permanent, défaits des institutions et des égos, qui conduiraient en permanence des argumentations et contre-argumentations logiques pour transformer un état de l'art en un *commons* scientifique.

Mais quel serait dès lors la source d'énergie d'une telle infrastructure ? Les chercheurs essayent d'être convaincants. On peut réussir un test de Turing en réalisant une projection convaincante ; ou en « sapant », en « sabotant », la capacité d'appréciation de l'audience. Une machine qui n'est pas intelligente peut ainsi générer un effet le plus proche de l'aspiration humaine à reconnaître ou vouloir croire ; les concurrents des épreuves annuelles des tests de Turing, consistant à prouver qu'une machine peut faire croire à un humain qu'elle est humaine, sont habitués de ce genre de stratagèmes. Elle fait, ou elle dit, ce qui est le plus proche de l'attente de la cible.

Le *deep learning* procède de la même manière. Sa performance d'imitation devient une asymptote à la réalité au fur et à mesure qu'on sature son apprentissage. Par exemple, les moteurs de traduction fondés sur l'apprentissage profond excellent dans l'imitation d'un style d'écriture[5] si bien que l'on peut facilement imaginer que

[5] Pour une démonstration d'un moteur de traduction fondé sur de l'apprentissage profond : www.deepl.com

les machines intelligentes de *proof-reading* et les agents conversationnels arriveront sans doute à produire automatiquement des textes au plus près des attentes de la communauté éditoriale, du *peer-review* d'*Administrative Science Quarterly* (ASQ) ou du *Journal of Management Studies* (JMS) avant 2030.

L'apprentissage adverse (*generative adversarial learning*), qui consiste à forcer des réseaux de neurones convolutionistes à apprendre contre eux-mêmes, peut sans doute offrir une substitution pertinente à l'intuition ; si au lieu d'entreprendre un apprentissage stationnaire, on force cet apprentissage sur une introspection non stationnaire (c'est-à-dire dans la recherche d'un modèle en dehors des tendances reconnaissables ou apprises qui expliquent mieux les données). En bref, on peut apprendre à un réseau, à un modèle, à ne s'intéresser qu'à ce qui le surprend ; et lui offrir une opposition dialectique, avec tout ce qui reçoit le plus grand score d'agrément scientifique.

Le problème est qu'il n'existe pas réellement de *ground truth* (de mesure étalon réelle) pour cette forme de théorisation automatique ; c'est-à-dire que l'établissement de la preuve passera toujours par une dimension politique : par la traduction d'une connaissance métier, par l'établissement et le contrôle de puits de données. On peut argumenter que l'aléa soit plus performant que le travail de la communauté scientifique : un apprentissage profond et non stationnaire cherchera tout modèle théorique, à partir d'un existant de modèles pouvant être combinés, permettant au mieux d'expliquer un phénomène. Il est à parier que cela sera sans doute moins caricatural que les productions qui s'enrobent de contre-intuitions pour se rendre intéressantes (Davis, 1971) ; ou qui sont produites par de véritables conglomérats du *proof-reading* et de vassalités éditoriales, cherchant à plaire à l'éditeur d'ASQ, du SMJ, ou de JMS, plus qu'à créer une révolution théorique.

L'autre problème d'une théorisation non stationnaire, non supervisée et autonome ; c'est-à-dire générée uniquement par des machines, est qu'elle risque de systématiser un effet « ELIZA » dans la production scientifique. L'effet « ELIZA » renvoie au nom de l'agent conversationnel de Joseph Weizenbaum (1966) qui était programmé pour imiter un psychothérapeute rogérien, mais qui abusait de techniques de relance placebo (« Qu'est-ce que X te suggère ? » ; « Je ne suis pas sûr de bien comprendre ») ; mais cela est déjà le cas sans apprentissage artificiel.

Le tableau 1 résume les quatre formes d'usage de l'IA pour la recherche identifiées dans cet article. Le premier quadrant (nord-ouest) est celui du paradigme

historique : des démarches stationnaires ; des révolutions scientifiques fondées sur des remises en cause des agréments considérés comme acquis ; une logique dominante hypothético-déductive. Sans surprise, l'intelligence artificielle y joue un rôle de productivité de la donnée (« data drives learning »). Ce quadrant pourrait s'intituler « le nouveau management scientifique » tant il en épouse l'idéologie et la téléologie. Sur le plan des implications managériales, le seul effet prévisible est la disparition – définitive – des *middle managers* dont le rôle contemporain n'a été que d'exercer une coercition servile à partir d'indicateurs clés de performance. N'importe quel apprentissage profond non supervisé peut en effet faire mieux qu'un humain en la matière.

Le quadrant sud-ouest est celui de l'apprentissage artificiel inductif formel ; ce qui est réalisé aujourd'hui par les réseaux convolutionistes ou l'apprentissage adverse. Ces approches permettent de dépasser non pas nos limites cognitives individuelles (le cerveau humain restant un calculateur à forte performance) ; mais la faiblesse de notre capacité individuelle ou collective d'imprégnation. Un chercheur confirmé maîtrise entre 500 et 1 000 références théoriques (articles qu'il ou elle a lu et est capable d'articuler aux 500/1 000 autres). Plus il ou elle vieillit, plus il ou elle va se réfugier dans des répertoires de réponses figées, ou possédera-t-il d'une plus grande capacité personnelle à tordre la réalité pour la faire entrer dans l'étroite fenêtre de sa culture scientifique personnelle. L'exploration machine inductive offre ici un réel espoir pour le renouvellement de toutes les sciences sociales ou sciences exactes à moyen terme ; dans la mesure où – dans une perspective stationnaire – un ensemble algorithmique peut sublimer une communauté scientifique, en corrigeant les travers et les limites de ses processus d'évaluation et d'adoption des nouvelles idées scientifiques.

**Tableau 1**
**Formes d'apprentissages machine et théorisation**

| | **Apprentissage stationnaire (explorer les variations de modèles connus)** | **Apprentissage non stationnaire (apprendre des modèles, pas des variables)** |
|---|---|---|
| **Exploration déductive** | – Améliorer l'efficacité du traitement de très larges données (Orion Spur, découverte d'exoplanètes, classification de galaxies, séquences ADN d'AlphaFold de DeepMind).<br>– IA *passive* ou *productive*, homomorphique aux corpus de connaissances établis, y compris dans l'institutionnalisation de la diffusion.<br>– Le nouveau management scientifique | – Amélioration des échantillonnages par le contrôle *ex ante* ou *ex post* de leur non-stationnarité (extraction de caractéristiques)<br>– Modélisation de phénomènes dont la non-stationnarité est intrinsèque (sujets aux dérives, systèmes complexes non déterministes, dont les propriétés probabilistes changent avec le temps : sciences cognitives, marchés financiers, biologie moléculaire, séquençage ADN, etc.) |
| **Exploration inductive** | – L'apprentissage profond et/ou adverse permet de rendre possible des sauts inductifs irréalisables avec la seule intuition humaine ; mais la pertinence de l'espace vectoriel sélectionné par la machine est aléatoire (et peu précis)<br>– Détection d'anomalies ou d'événements rares et inhabituels dans de larges sets de données.<br>– « Agents conversationnels » permettant de conduire une discussion scientifique (verbale), de la rédiger, et de dialoguer avec le chercheur (lève les barrières disciplinaires) | – Découverte de relations *inconnues* entre des séries de variables ; production autonome et non supervisée de *modèles* à partir de données brutes<br>– Détection des faux positifs et des faux négatifs, en essayant de générer une infinité de modèles pouvant expliquer un phénomène. Le phénomène restant inexpliqué est un faux positif.<br>– Apprentissage adverse de *machine peer review* (l'algorithme fait la recension en cherchant à générer un modèle qui explique mieux que les modèles soumis) |

Les deux derniers quadrants sont les plus passionnants (nord-est et sud-est) car ils sont à même de révolutionner la recherche scientifique en général, et le management en particulier. La plupart des champs scientifiques, soumis à des évaluations quantitatives de la performance, forcent les chercheurs à sacrifier le labeur de l'abstraction et de la créativité, par celui de la conformation et de la maximisation de l'espérance de publication (le fameux *publish or perish*, qui est asymptotiquement littéral). Le premier quadrant (nord-est) est déjà à l'œuvre. Les modèles d'apprentissages artificiels les plus récents sont tout à fait capables, dans un paradigme déductif, de faire émerger des modèles (des abstractions, des découvertes de relations insoupçonnées) à partir de corpus non stationnaires (au sein desquels il est impossible d'établir une tendance ou de figer un modèle *a priori*). Ces développements qui concernent l'astronomie, la physique quantique, le séquençage de l'ADN, n'ont pas, pour l'instant, de monétisations évidentes, et n'ont pas encore réussi à modifier les habitus de la recherche scientifique.

Le dernier quadrant (sud-est) est celui d'une intelligence artificielle apprenant des modèles, pas des données. Il combine une exploration inductive qui reprend littéralement à son compte les intuitions de Boole, Turing, Ivakhnenko et Blaug (ou encore de Starbuck, 1983) : les êtres humains recherchent généralement des problèmes qui peuvent expliquer leurs solutions, et rarement l'inverse ; si bien que la cognition *machine*, pour se rapprocher de l'exceptionnelle versatilité et performance de la cognition humaine, devrait s'inspirer d'une conception qui découvre de manière

autonome l'ensemble des modèles pouvant expliquer une donnée ; la vérification se réalisant par l'existence d'un modèle explicatif, plus ou moins proche en termes de probabilités (Boole) de la conclusion la plus vraisemblable (Weick). À l'instar du cerveau humain, et telle fut l'intuition de la cybernétique à sa fondation, une machine doit être motivée non pas par le systématisme de ses requêtes mais par sa capacité à se surprendre elle-même, à savoir ce qui est incongru, à se passionner pour les événements qui résistent à son pouvoir explicatif. Dès lors, les machines pourront sans doute théoriser à l'égal des hommes, avec quelques avantages (l'institutionnalisation et la cooptation algorithmiques prendront du temps), et quelques inconvénients (il faudra modéliser la persévérance dans l'échec ; une grande qualité humaine).

## Conclusion

Nous passons beaucoup de temps à tâtonner, accrochés à l'espoir que nous avons trouvé une question intéressante (Davis, 1971) ; qui puisse susciter l'intérêt d'un comité de lecture ; devenir une publication ; avoir ensuite une réelle existence : être citée ou être discutée. La façon dont nous produisons nos recherches demande des efforts titanesques et constitue une forme archaïque de labeur humain. Nous naviguons dans des quantités exponentielles de références, avec pour seul outil d'excavation un moteur sémantique très primaire, composé de nos yeux, de notre heuristique biaisée et imparfaite et d'un clavier. Nous nous épuisons à lire des quantités d'articles qui ne répondent pas à nos questions ; qui ne sont pas reproductibles ; dont les résultats sont laissés « à notre libre interprétation ». Nous composons des hypothèses qui nous semblent prometteuses ; nous instrumentons ; nous échouons ; nous soumettons, resoumettons.

Ce processus est doublement stationnaire ; d'une part, dans la méthodologie individualiste de production (*it's all about us*), parce que nous assumons que nous devons trouver un modèle (une récurrence, un motif, un schéma) que nous soumettons à l'épreuve des tests empiriques ; et d'autre part, parce que le moteur d'acceptation de nos propositions théoriques (nos modèles, nos articles) est une hybridation de cooptation institutionnelle et de recensions en double aveugle dont la fiabilité et la capacité réelle d'exploration sont douteuses.

Les premiers rituels d'évaluation en double aveugle furent créés en Grande Bretagne au XIX$^{e}$ siècle ; mais le rôle des évaluateurs n'était pas de sanctionner les articles, mais d'accompagner leur amélioration (Csiszar, 2016). Les rapports rédigés

par les académiciens étaient destinés à être débattus en public ; un système qui perdura jusqu'à la moitié du XIX$^{e}$ siècle, avant, que peu à peu, le secret soit réclamé sur les rapports, puis les débats eux-mêmes. L'anonymat et le formalisme de ces rapports ne virent finalement le jour qu'au début du XX$^{e}$ siècle. La plupart des journaux scientifiques n'avaient ainsi aucun processus d'évaluation formel avant l'après Seconde Guerre mondiale. Zyman (1968) suggère que cette formalisation n'avait pas beaucoup avoir avec la matière scientifique, mais résultait d'une solidification du métier de « scientifique » dans la société. Si un employé de bureau devait répondre à ses supérieurs, comment pouvait-on s'assurer du même égard envers un scientifique ? Les budgets qui furent engagés dans le cadre de la reconstruction, du Plan Marshall et de la création du bureau présidentiel pour la science aux États-Unis, dont Herbert Simon était un membre, jouèrent également un rôle décisif dans la formalisation des processus de *peer review*.

La recherche contemporaine en sciences sociales est l'objet de trois formes de critiques récurrentes : son incapacité à proposer des questions intéressantes (Davis, 1971 ; Bartunek *et al.*, 2006) ; son déclin dans la production de théorisations conceptuelles transformatives (Grandy et Mills, 2004 ; Hillman, 2011 ; Baumard, 2017 ; Starbuck et Baumard, 2009) ; et sa production exponentielle d'articles scientifiques dont les résultats ne peuvent être reproduits (Alvesson et Sandberg, 2013 ; Ioannidis, 2005). La recherche scientifique humaine n'a jamais été aussi abondante. Le seuil des 50 millions d'articles scientifiques publiés dans le monde fut franchi en 2010. Depuis la création du *Journal des savants* en France en 1665, la production humaine d'articles de recherche n'a cessé de croître pour atteindre un niveau de production annuelle de 2,5 millions de publications scientifiques (à comité de lecture), à un taux de croissance dépassant les 10 % depuis 2010, soit l'équivalent d'un doublement du volume de publications scientifiques mondiales tous les neuf ans (Bornmann et Mutz, 2015).

L'intelligence artificielle n'est pas une révolution de la donnée ; et les données ne seront jamais des théories. Les apprentissages profonds sont incapables de découverte. Qu'ils soient convolutionistes, ou qu'ils soient adverses, ils ne font que construire le chemin inductif entre des éléments épars qui sont peut-être inconnus à l'observateur, mais qui – toujours – sont eux-mêmes issus d'une réalité préexistante. La seule intelligence artificielle est celle qui est capable d'inventer un modèle explicatif, inconnu de l'homme, ignoré de l'état de l'art, pouvant expliquer une donnée ou une conclusion, sans le recours à la reconnaissance combinatoire.

Pour les sciences de l'organisation, cela implique de former à la recherche différemment. Il ne faut plus enseigner les canons méthodologiques, et attendre une soumission dévote à des dogmes quantitatifs ou qualitatifs, mais enseigner au contraire la modélisation non stationnaire, non supervisée ; en somme, se libérer de l'incarcération disciplinaire, et refaire des sciences de l'organisation une discipline cybernétique, comme l'avait désiré un de ses fondateurs, Herbert Simon. L'intelligence artificielle est chez elle dans ces sciences de l'organisation emprunteuses, fouineuses, indisciplinées, voleuses, qui ont réussi depuis le groupe RAND du Carnegie Institute of Technologie à emprunter à toutes les disciplines possibles et imaginables. Que peut-on dès lors espérer de plus enivrant que ces débats vifs avec des agents conversationnels, qui porteront en eux aussi bien la vivacité d'esprit d'un Turing, le mysticisme mécaniste d'un Ivakhnenko (1970), tout en étant capable de produire des modèles qu'une pensée humaine courante n'aurait jamais su concevoir. La vraie question reste, et a toujours été : *What's next ?*

Mis en ligne sur Cairn.info le 04/03/2020
https://doi.org/10.3166/rfg.2020.00409