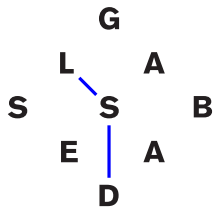


JOURNAL > SITE 1 : PORTE LOGIQUE, POLITIQUES DE L'ESPRIT ARTEFACTUEL | 2017

Des machines qui morphent la logique : les réseaux de neurones et l'automatisation déformée de l'intelligence comme inférence statistique

Matteo Pasquinelli

> *Les perceptrons [réseaux de neurones artificiels] ne sont pas destinés à servir de copies détaillées d'un quelconque système nerveux réel. Ce sont des réseaux simplifiés, conçus pour permettre l'étude des relations réglées entre l'organisation d'un système nerveux, l'organisation de son environnement et les performances « psychologiques » dont le réseau est capable. Les perceptrons pourraient correspondre en réalité à des parties de réseaux plus étendus dans les systèmes biologiques. [...] Plus vraisemblablement, ils représentent des simplifications extrêmes du système nerveux central, dans lesquelles certaines propriétés sont exagérées et d'autres supprimées.* (*Frank Rosenblatt, 1961*[1])
>
> *Il n'existe pas d'algorithme pour créer des métaphores, et une métaphore ne peut non plus être produite en fournissant des instructions précises à un ordinateur, quel que soit le volume d'information organisée qu'on y entre.* (*Umberto Eco, 1986*[2])

Dans la presse grand public, mais aussi dans les cercles artistiques et philosophiques, l'expression « intelligence artificielle » fait souvent office de talisman alchimique. Mais son fonctionnement est rarement expliqué. À ce jour, le paradigme hégémonique (paradigme qui possède aussi une importance cruciale dans l'automatisation du travail) ne repose pas sur la GOFAI (la « *Good Old-Fashioned Artificial Intelligence* » [la bonne vieille intelligence artificielle], qui n'est jamais parvenue à automatiser la *déduction symbolique*), mais sur les réseaux neuronaux conçus par Frank Rosenblatt en 1958 pour

automatiser l'*induction statistique*. Cet article met en lumière le rôle joué par les portes logiques dans l'architecture distribuée des réseaux neuronaux, où une boucle de contrôle généralisée affecte chaque nœud computationnel afin d'effectuer une reconnaissance de motifs. Dans l'architecture distribuée et adaptative des portes logiques, la logique n'est pas appliquée à l'information selon un schéma *top-down* ; au contraire, *l'information se transforme en logique*, ou, en d'autres termes, une représentation du monde devient une nouvelle fonction dans la même description du monde. Je propose cette formulation élémentaire comme définition plus précise de l'*apprentissage* afin de contester la définition idéaliste de l'intelligence (artificielle). Si la reconnaissance de motifs fondée sur l'induction statistique constitue le descripteur le plus précis de ce que l'on appelle « intelligence artificielle » dans le langage populaire, il reste encore à comprendre complètement les effets déformants de l'induction statistique sur la perception collective, l'intelligence et la gouvernance (surapprentissage, apophénie, biais algorithmique, « *deep dreaming* », etc.).

Plus généralement, ce texte avancera l'hypothèse selon laquelle les nouvelles machines enrichissent et déstabilisent les catégories mathématiques et logiques qui ont contribué à leur conception. Une machine est toujours une machine de cognition, produit de l'intellect humain et composante indisciplinée des rouages de la cognition étendue. C'est grâce aux machines que l'intellect humain peut traverser, sur un mode matérialiste, de nouveaux paysages de la logique – c'est-à-dire sous l'influence, non de l'idéalisme, mais d'artefacts historiques. De même, par exemple, que le moteur thermique a suscité l'émergence de la science de la thermodynamique (et non l'inverse), de même, on peut attendre des machines computationnelles qu'elles jettent un éclairage neuf sur la philosophie de l'esprit et la logique elle-même. En inventant l'idée de machine computationnelle universelle, Alan Turing entendait calculer toutes les fonctions possibles à partir de la mécanisation la plus simple. L'efficacité du calculateur universel cristallisa chez Turing le projet alchimique d'une automatisation de l'intelligence humaine. Cependant, il y aurait un formidable paradoxe à ce que la machine de Turing, expérience de pensée destinée à démontrer l'incomplétude de la mathématique, puisse servir l'ambition de décrire un paradigme exhaustif de l'intelligence (car c'est ainsi que l'on comprend souvent le test de Turing).

Une unité d'information est une unité logique de décision

Plutôt que de réitérer la GOFAI, autrement dit, l'application *top-down* de la logique à l'information récupérée dans le monde, ce texte tente d'exposer la transmutation de *l'information externe en logique interne* dans la mécanisation des réseaux neuronaux. Au sein des réseaux neuronaux (mais aussi selon le cadre de la cybernétique classique), l'information devient contrôle : en d'autres termes, un input numérique reçu du monde devient une fonction de contrôle de ce même monde. Ou, pour employer un langage plus philosophique, cela signifie qu'une représentation du monde (information) devient une nouvelle règle dans le même monde (fonction), mais avec un fort degré d'approximation statistique. *L'information devient logique* : voilà une formulation très grossière de l'intelligence, mais qui tente de présenter l'ouverture au monde comme processus continu d'apprentissage.

La transformation de l'information en fonctions supérieures se retrouve sans doute à différents stades de l'histoire des machines intelligentes ; ce texte ne s'attachera qu'à la définition originelle de l'information et des boucles de rétroaction, puis il analysera leurs ramifications dans les réseaux de neurones. La métamorphose d'une boucle d'information en formes supérieures de connaissance du monde fut l'objet de la cybernétique de second ordre au cours des années 1970, mais les réseaux neuronaux de Rosenblatt, à la fin des années des années 1950, en étaient déjà un exemple[3]. Pour comprendre comment les réseaux neuronaux transforment l'information en logique, il pourrait donc être utile de déconstruire la réception traditionnelle des concepts d'information et de rétroaction de l'information. On a coutume de critiquer Claude Shannon pour avoir réduit l'information à une mesure mathématique liée au bruit dû au canal[4]. De façon plus intéressante, à la même époque, Norbert Wiener a défini l'information comme *décision*.

> *Qu'est-ce que l'information, et comment la mesure-t-on ? L'une des formes les plus simples et les plus unitaires d'information réside dans l'enregistrement d'un choix entre deux possibilités simples, dotées d'un même degré de probabilité, et dont l'une ou l'autre va se produire – par exemple, un choix entre pile et face lorsque l'on lance une pièce. Nous appellerons* décision *un choix unique de ce type*[5].

Si chaque unité d'information est une unité de décision, alors l'information recèle une doctrine atomique du contrôle. Si l'information est décision, tout bit d'information est une petite pièce de la logique de contrôle. Bateson ajoutera, dans une formule célèbre, que « l'information est une différence qui fait une différence », préparant ainsi la cybernétique à des ordres supérieurs d'organisation[6]. En fait, la cybernétique de second

ordre est venue rompre l'emprise exercée par la boucle de rétroaction négative et avec la cybernétique des débuts, qui tenait absolument à maintenir constamment à l'équilibre les systèmes biologiques, techniques et sociaux. Une boucle de rétroaction négative se définit comme une boucle d'information qui est graduellement ajustée pour adapter un système à son environnement (en régulant sa température, sa consommation d'énergie, etc.). Une boucle d'information positive est, à l'inverse, une boucle qui devient incontrôlable et entraîne un système loin de l'équilibre. La cybernétique de second ordre a noté que les systèmes éloignés de l'équilibre rendent possible la naissance de nouvelles structures, habitudes et idées (le prix Nobel Ilya Prigogine a montré que des formes d'auto-organisation se produisent aussi dans des états turbulents et chaotiques)[7]. Certes, dans la formulation élémentaire de la première cybernétique, on pouvait déjà comprendre la boucle de rétroaction comme un modèle d'information qui se transforme en logique, qui morphe la logique elle-même pour inventer de nouvelles règles et habitudes. Mais ce que semble introduire la cybernétique de second ordre, c'est l'idée que la « pression » excessive du monde extérieur force la logique machinique à muter.

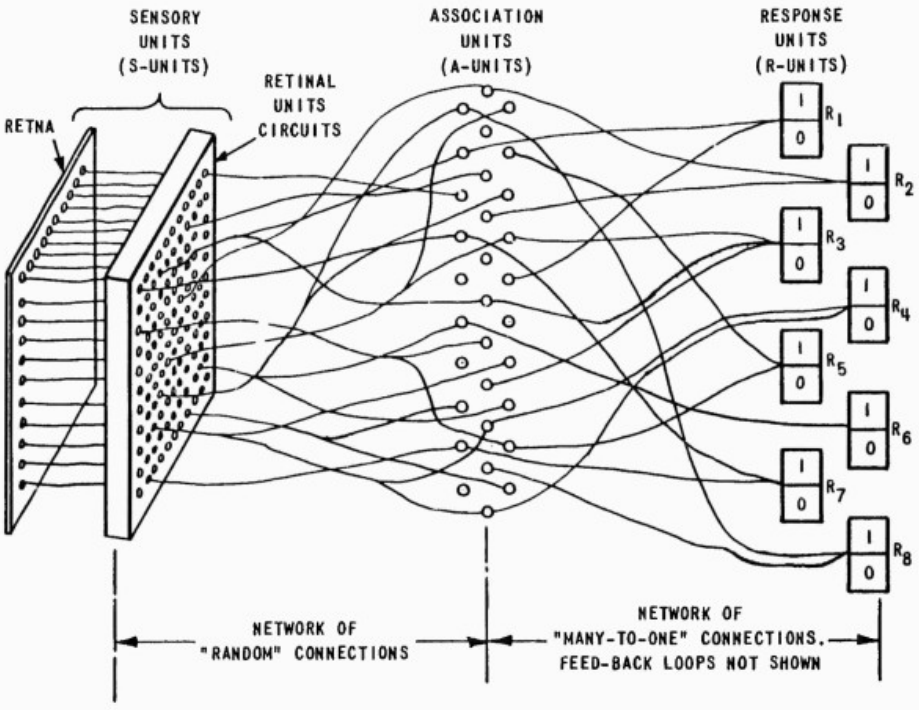


Figure 1 ORGANIZATION OF THE MARK 1 PERCEPTRON

Diagramme de l'organisation du Mark 1 Perceptron. La source comprenant la boucle de rétroaction n'est pas représentée. Source: Frank Rosenblatt, Mark I Perceptron Operators' Manual. Buffalo, NY: Cornell Aeronautical Laboratory, 1960.

Frank Rosenblatt et l'invention du Perceptron

Dans les généalogies multiples de l'évolution de l'intelligence artificielle, ce texte ne retiendra que la confrontation entre deux anciens camarades de la High School of Science du Bronx, Marvin Minsky, fondateur de l'Artificial Intelligence Lab du MIT, et Frank Rosenblatt, inventeur du premier réseau neuronal opérationnel, le Perceptron. On réduit souvent l'opposition entre Minsky et Rosenblatt à une dispute entre un paradigme *top-down* fondé sur des règles (l'IA symbolique) et le calcul parallèle et distribué (le connexionnisme). Dans le second modèle, la machine n'incarne pas dès le départ un algorithme pleinement intelligent ; au contraire, elle *apprend* de l'environnement et, progressivement, elle devient en partie « intelligente ». En termes logiques, il y a ici une tension entre *déduction symbolique* et *induction statistique*[8].
En 1951, Minsky développa le premier réseau de neurones artificiels, le SNARC, projet qu'il abandonna ensuite, convaincu que les réseaux neuronaux nécessitaient une trop grande puissance de calcul[9]. En 1957, Rosenblatt décrivit le premier réseau neuronal fonctionnel dans un rapport destiné au laboratoire d'aéronautique de Cornell et intitulé « The Perceptron: A Perceiving and Recognizing Automaton ». Comme Minsky, Rosenblatt esquissait son propre réseau neuronal, mais en donnant une structure *bottom-up* et distribuée à l'idée de neurone artificiel développée par Warren McCulloch et Walter Pitts, elle-même inspirée par les neurones oculaires[10]. La première machine neuronale, le Mark 1 Perceptron, était en réalité une machine de vision[11].

> *Le réquisit premier d'un tel système est qu'il soit capable de reconnaître des schémas complexes d'information qui sont similaires sur le plan phénoménal [...], un processus qui correspond aux phénomènes d'« association » et de « généralisation de stimulus ». Le système doit reconnaître le « même » objet dans différentes orientations, tailles, couleurs ou transformations et sur une variété de fonds différents. [Il] devrait être possible de construire un système électronique ou électromécanique qui apprendra à reconnaître les similitudes ou les identités entre les schémas d'information optique, électrique ou tonale, d'une façon qui puisse être étroitement analogue aux processus perceptifs d'un cerveau biologique. Le système proposé repose, quant à son fonctionnement, sur des principes probabilistes plutôt que déterministes et il acquiert sa fiabilité grâce aux propriétés des mesures statistiques obtenues à partir de grandes populations d'éléments*[12].

Il faut préciser que le Perceptron n'était pas une machine destinée à reconnaître des formes simples, comme des lettres (la reconnaissance optique de caractères existait déjà à l'époque), mais une machine qui pouvait *apprendre* à reconnaître des formes en calculant un seul fichier statistique au lieu d'en sauvegarder un grand nombre dans sa mémoire. Au-delà de la reconnaissance d'image, Rosenblatt faisait ce pronostic prophétique : « On attend *in fine* des appareils de ce type qu'ils soient capables de former des concepts, de traduire des langues, de collecter des renseignements militaires et de résoudre des problèmes grâce à la logique inductive[13]. »
En 1961, Rosenblatt a publié *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanism*, qui a influencé jusqu'à ce jour la computation neuronale (par exemple, l'expression « Perceptron multicouche » est déjà là). L'ouvrage tire parti des découvertes psychologiques et neurologiques sur la neuroplasticité et les applique à la conception des réseaux neuronaux. Le perceptron était un modèle artefactuel du cerveau destiné à expliquer certains de ses mécanismes sans être pris pour le cerveau lui-même (en fait, les réseaux neuronaux imitaient les neurones de l'œil et non du cerveau, à une époque où l'on ne savait pas encore comment le cortex visuel traite les inputs visuels). Rosenblatt soulignait que les réseaux neuronaux artificiels étaient à la fois une simplification et une exagération des systèmes nerveux et que cette approximation (autrement dit, la reconnaissance des limites d'une approche en termes de modèle) devait servir de guide à la philosophie de l'esprit (artefactuel). Enfin, Rosenblatt défendait la *neurodynamique* comme discipline, contre la vogue de l'intelligence artificielle.

> *Le programme perceptron ne vise* pas *d'abord à inventer des appareils d'« intelligence artificielle », mais plutôt à examiner les structures physiques et les principes neurodynamiques qui sous-tendent l'« intelligence artificielle ». Un perceptron est d'abord et avant tout un modèle du cerveau, non une invention destinée à la reconnaissance de schémas. En tant que modèle du cerveau, son utilité réside dans le fait qu'il nous permet de déterminer les conditions physiques de l'émergence de diverses propriétés psychologiques. Il ne s'agit en aucun cas d'un modèle « complet », et nous avons pleinement conscience des simplifications qu'il implique par rapport aux systèmes biologiques ; mais il s'agit au moins d'un modèle analysable*[14].

En 1969, Marvin Minsky et Seymour Papert attaquèrent, dans un livre intitulé *Perceptrons*, le modèle de réseau neuronal élaboré par Rosenblatt, en affirmant à tort qu'un perceptron (bien que monocouche) ne pouvait apprendre la fonction XOR et résoudre des classifications dans des dimensions supérieures. Cet ouvrage récalcitrant eut un impact dévastateur, en partie aussi à cause de la mort prématurée de Rosenblatt

en 1971 : pendant plusieurs années, la recherche sur les réseaux neuronaux n'eut plus de financements. Au lieu de parler, comme on a coutume de le faire, du premier « hiver de l'intelligence artificielle », on devrait parler de l'« hiver des réseaux neuronaux ». Celui-ci dura jusqu'en 1986, lorsque l'ouvrage *Parallel Distributed Processing* démontra que les perceptrons (multicouches) peuvent bel et bien apprendre des fonctions logiques complexes[15]. Un demi-siècle et d'innombrables neurones plus tard, n'en déplaise à Minsky, Papert et aux fondamentalistes de l'IA symbolique, les perceptrons multicouches sont capables d'une meilleure reconnaissance d'image que les humains et constituent le cœur des systèmes de *Deep Learning*, tels que la traduction automatique et les voitures autonomes[16].

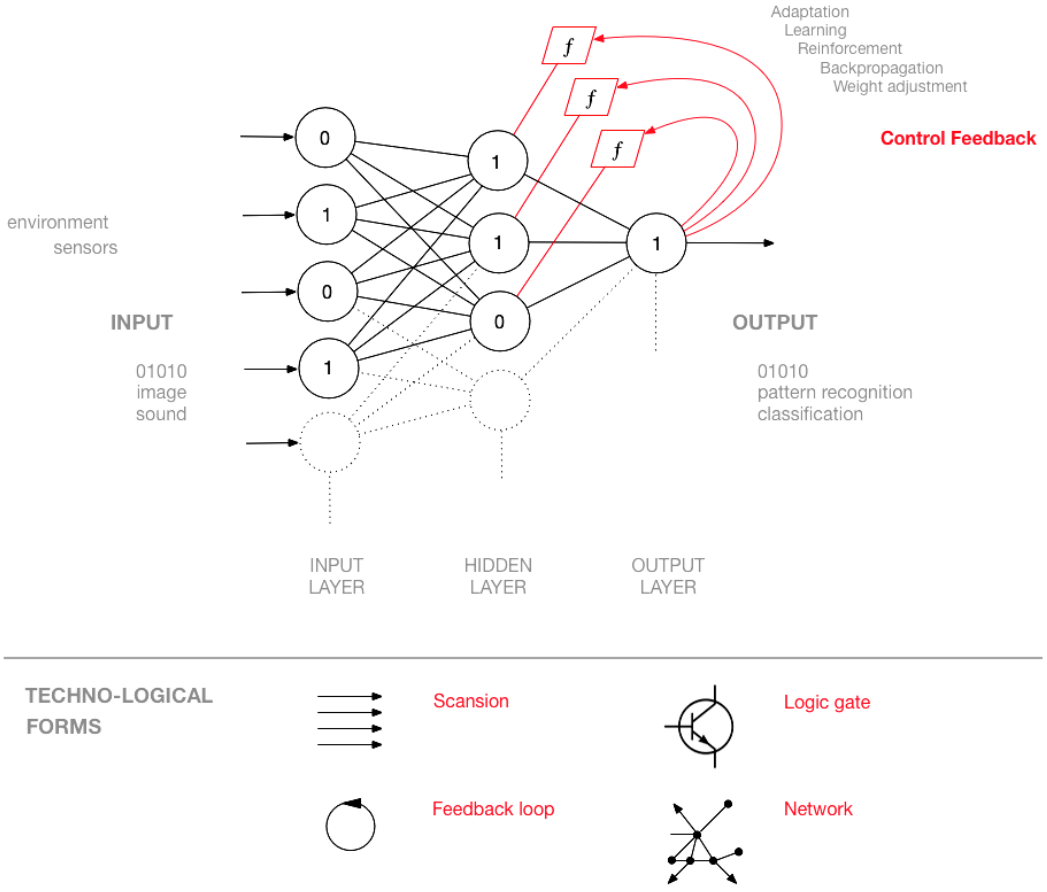


Diagramme d'un réseau neuronal simple montrant des boucles de rétroaction. Matteo Pasquinelli, HfG Karlsruhe. Voir: www.academia.edu/33205589

Anatomie d'un réseau neuronal

Sur le plan de l'archéologie des media, on peut décrire l'invention des réseaux neuronaux comme la composition de quatre formes techno-logiques : la scansion (discrétisation ou numérisation d'inputs analogiques), la porte logique (qui peut prendre la forme du potentiomètre, de la soupape, du transistor, etc.), la boucle de rétroaction (idée fondamentale de la cybernétique) et le réseau (inspiré ici par l'agencement des neurones et des synapses). Néanmoins, un réseau de neurones a pour finalité de calculer une construction statistico-topologique plus complexe que la disposition de ces formes. La fonction d'un réseau neuronal est d'enregistrer des motifs similaires d'inputs (jeux de données de formation) comme *état interne* de ses nœuds. Une fois qu'un état interne a été calculé (en d'autres termes, une fois que le réseau neuronal a été « formé » pour reconnaître un motif spécifique), cette construction statistique peut être intégrée dans des réseaux neuronaux possédant une structure identique et utilisée pour reconnaître des motifs dans de nouvelles données.

Les portes logiques, qui font d'ordinaire partie des structures linéaires de calcul, acquièrent, dans le calcul parallèle des réseaux neuronaux, de nouvelles propriétés. En ce sens, Rosenblatt est sans doute l'un des premiers à avoir décrit l'intelligence des machines comme propriété émergente : « Il est significatif qu'il n'ait jamais été démontré que les éléments individuels, ou les cellules, d'un réseau nerveux possèdent des fonctions spécifiquement psychologiques, telles que la "mémoire", la "conscience" ou l'"intelligence". Par conséquent, on peut présumer que de telles propriétés résident dans l'organisation et le fonctionnement du réseau dans son ensemble, plutôt que dans ses parties élémentaires[17]. » Or les réseaux de neurones ne sont pas horizontaux mais hiérarchisés (ils possèdent plusieurs couches).

Le réseau neuronal se compose de trois types de couches de neurones : la couche d'entrée, les couches cachées (qui peuvent être nombreuses, d'où l'expression « Deep Learning ») et la couche de sortie. Depuis le premier perceptron (chose qui révèle l'influence du paradigme visuel), la couche d'entrée est souvent appelée « rétine », même si elle ne calcule pas de données visuelles. Les neurones de la première couche sont connectés aux neurones de la suivante, selon un flux d'information dans lequel un input complexe est encodé pour correspondre à un output donné. La structure qui en émerge n'est pas exactement un réseau (ou un rhizome) mais un *réseau arborescent* qui se développe en tant que *cône hiérarchique* où l'information est acheminée et distillée dans des formes supérieures d'abstraction[18].

Chaque neurone du réseau est un nœud de transmission mais aussi un nœud computationnel – porte informationnelle et porte logique. Chaque nœud possède donc un double rôle : transmettre l'information et appliquer la logique. Le réseau neuronal « apprend » à mesure que le mauvais output est réorienté pour ajuster l'erreur de chaque nœud de computation jusqu'à ce que soit obtenu l'output désiré. Les réseaux neuronaux sont bien plus complexes que les systèmes cybernétiques traditionnels, puisqu'ils exemplifient une boucle de rétroaction généralisée, qui affecte une multitude de nœuds computationnels. En ce sens, le réseau de neurones constitue l'architecture de calcul la plus adaptative qui ait été conçue en matière d'apprentissage automatique.
La rétroaction généralisée affecte la *fonction* de chaque nœud ou neurone, autrement dit, la manière dont un nœud calcule (sa « pondération »). La rétroaction qui contrôle le calcul de chaque nœud (et qui prend des noms divers : « ajustement de pondération », « rétropropagation de l'erreur », etc.) peut être une équation, un algorithme, voire un opérateur humain. Dans un cas spécifique de réseau neuronal, lorsque l'on modifie le seuil d'un nœud, la rétroaction de contrôle peut transformer une porte OU en porte ET, par exemple – ce qui signifie que la rétroaction de contrôle change la manière dont un nœud « pense »[19]. Les portes logiques des réseaux neuronaux calculent de l'information afin de modifier la manière dont elles calculeront de l'information future. Ainsi, *l'information affecte la logique*. Aujourd'hui, le principal souci commercial des grandes compagnies des technologies de l'information est de découvrir la formule la plus efficace de rétroaction de contrôle neuronal.
Plus spécifiquement, le réseau de neurones apprend à reconnaître une image en enregistrant les dépendances ou les relations entre les pixels et en composant statistiquement une représentation interne. Dans une photo de pomme, par exemple, un pixel rouge peut être entouré par d'autres pixels rouges 80 % du temps, etc. De cette façon aussi, des relations inhabituelles peuvent être combinées dans des caractéristiques graphiques plus complexes (bords, lignes, courbes, etc.). Comme une pomme doit pouvoir être reconnue sous différents angles, ce n'est jamais une image réelle qui est mémorisée, mais seulement ses dépendances statistiques. Le graphe statistique des dépendances est enregistré comme une représentation interne multidimensionnelle et qui sera ensuite associée à un output lisible par les humains (le mot « pomme »). Ce modèle de formation est appelé « apprentissage supervisé », puisqu'un être humain décide du caractère correct ou incorrect de chaque output. On parle d'apprentissage non supervisé lorsque le réseau neuronal doit découvrir les schémas de dépendance les plus

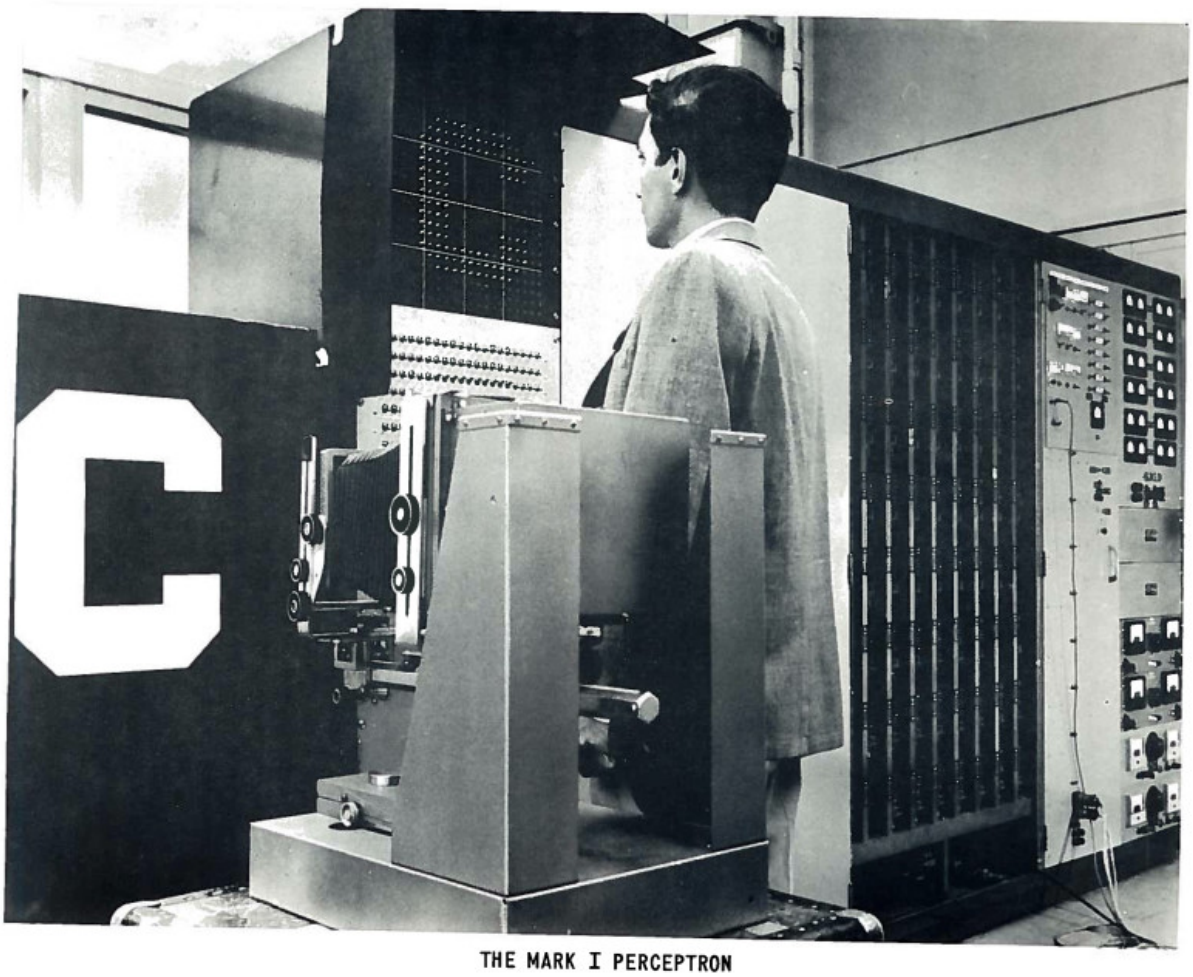
communs dans un jeu de données de formation sans suivre une classification préalable (à partir d'un jeu de données d'images de chats, il extraira les caractéristiques d'un chat générique).
Dépendances et motifs peuvent être repérés dans les types de données les plus divers : si les données visuelles sont les plus compréhensibles intuitivement, les mêmes procédures s'appliquent aux données sociales, médicales et économiques, par exemple. À l'évidence, les techniques actuelles de l'intelligence artificielle relèvent davantage d'une forme sophistiquée de reconnaissance de motifs que de l'intelligence, si l'on conçoit l'intelligence comme *la découverte et l'invention de nouvelles règles*. Ou, pour le dire dans des termes logiques précis, c'est une forme d'*induction statistique* que les réseaux neuronaux calculent. Bien sûr, une forme aussi extraordinaire d'inférence automatisée peut être un précieux auxiliaire de la créativité et de la science humaines (et constitue la meilleure approximation de ce que l'on appelle l'abduction faible de Peirce), mais elle ne représente pas en soi l'automatisation de l'intelligence en tant qu'invention, précisément parce qu'elle reste confinée à des catégories « trop humaines »[20].

Une computation humaine, trop humaine

Peirce disait que « l'homme est un signe externe[21] ». Si cette intuition a incité les philosophes à souligner que l'esprit humain est un projet artefactuel qui se prolonge dans la technologie, l'imbrication effective de l'esprit humain et des machines externes de cognition n'a que rarement bénéficié d'illustrations *empiriques*. Cela a conduit à des postures simplistes, consistant à brandir les idées d'intelligence artificielle générale ou de superintelligence en talismans alchimiques du posthumanisme mais ne fournissant guère d'explications sur les rouages internes et les postulats du calcul. Un aspect fascinant du calcul neuronal réside dans la manière dont, loin de les dépasser dans des formes autonomes, il amplifie les catégories de la connaissance humaine. Contrairement à la conception naïve de l'autonomie de l'intelligence artificielle, de nombreux éléments de l'architecture des réseaux neuronaux demeurent profondément affectés par l'intervention humaine. Si l'on veut comprendre en quoi le calcul neuronal se prolonge dans l'« inhumain », il faut saisir en quoi il demeure « trop humain ». Le rôle de l'humain (et aussi le site du pouvoir) est clairement visible dans (1) la conception du jeu de données de formation et ses catégories, (2) la technique de correction d'erreur et (3) la classification de l'output désiré. Par manque de place, je me limiterai ici au premier point.

La conception du jeu de données de formation constitue la composante la plus critique et la plus vulnérable de l'architecture des réseaux neuronaux. Le réseau de neurones est formé pour reconnaître des motifs dans des données passées, dans l'espoir d'élargir cette capacité à des données futures. Mais, comme cela s'est déjà produit à de nombreuses reprises, si les données de formation présentent un biais de race, de genre et de classe, les réseaux neuronaux ne feront que le refléter, l'amplifier et le déformer. Les systèmes de reconnaissance faciale formés à partir de bases de données de visages de personnes blanches ont lamentablement échoué à reconnaître des Noirs comme des humains. Ce problème se nomme *surapprentissage* : un réseau neuronal disposant d'une grande puissance de calcul manifestera une tendance à trop apprendre, c'est-à-dire à se focaliser sur un motif très spécifique. Il sera par conséquent nécessaire d'éliminer certains de ses résultats afin d'assouplir sa capacité de reconnaissance. On peut considérer que le cas de l'*apophénie* est similaire au surapprentissage : par exemple, dans les paysages psychédéliques produits par le programme DeepDream de Google, les réseaux neuronaux « voient » des motifs qui n'y sont pas, ou, mieux encore, génèrent des motifs sur fond de bruit. Le surapparentissage et l'apophénie sont deux exemples des limites intrinsèques de la computation neuronale : ils montrent qu'au lieu de contribuer à révéler de nouvelles corrélations, les réseaux neuronaux peuvent entrer dans une spirale paranoïaque à cause des motifs qu'ils ont incorporés.

Le surapprentissage soulève une question plus fondamentale, relative à la constitution des jeux de données de formation : la limite des catégories au sein desquelles opère le réseau neuronal. La manière dont un jeu de données de formation représente un échantillon du monde délimite, en même temps, un univers clos. Quelle est la relation de cet univers de données clos avec l'extérieur ? On considère qu'un réseau neuronal est « formé » lorsqu'il est capable de généraliser ses résultats à des données inconnues avec une très faible marge d'erreur, mais cette généralisation est rendue possible par l'homogénéité entre les jeux de données de formation et les jeux de données de test. Se pose donc la question suivante : dans quelle mesure un réseau neuronal (et l'IA en général) est-il capable d'échapper à l'ontologie catégorielle dans laquelle il opère[22] ?



Mark 1 Perceptron. Source: Rosenblatt, Frank (1961) Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Buffalo, NY: Cornell Aeronautical Laboratory.

L'abduction de l'inconnu

D'ordinaire, un réseau neuronal calcule l'induction statistique à partir d'un jeu de données homogène : il extrapole des motifs compatibles avec la nature du jeu de données (un motif visuel à partir de données visuelles, par exemple). Mais si le jeu de données n'est pas homogène et contient des caractéristiques multidimensionnelles (pour prendre un exemple très simple, des données sociales décrivant l'âge, le genre, le revenu, le niveau d'éducation et de santé de la population, etc.), les réseaux neuronaux peuvent découvrir des motifs parmi des données que la cognition humaine n'a pas tendance à corréler. Même si les réseaux neuronaux montrent des corrélations que l'esprit humain ne peut prévoir, ils opèrent au sein de la grille implicite des postulats et des catégories (humains) qui se trouvent déjà dans le jeu de données de formation, et, en ce sens, ils ne peuvent effectuer le saut nécessaire à l'invention de catégories radicalement nouvelles.

La distinction effectuée par Charles S. Peirce entre déduction, induction et abduction (hypothèse) est le meilleur moyen de définir les limites et potentialités de l'intelligence des machines. Peirce a remarquablement vu que les formes logiques classiques de l'inférence – la déduction et l'induction – n'inventent jamais de nouvelles idées mais ne font que répéter des faits quantitatifs. Seule l'abduction (l'hypothèse) est capable de concevoir de nouvelles visions du monde et d'inventer de nouvelles règles.

> *L'induction n'accomplit qu'une seule chose : elle détermine la valeur d'une quantité. Elle part d'une théorie et mesure le degré de concordance de cette théorie avec les faits. Elle ne saurait jamais engendrer d'elle-même une quelconque idée. Il en va de même de la déduction. Toutes les idées de la science adviennent par le biais de l'Abduction. L'abduction consiste à étudier les faits et à concevoir une théorie pour les expliquer*[23].

Plus spécifiquement, la distinction entre abduction et induction peut éclairer la forme logique des réseaux de neurones, qui, depuis leur invention par Rosenblatt, sont conçus pour automatiser des formes complexes d'induction.

> *Par l'induction, nous concluons que les faits similaires aux faits observés sont vrais dans les cas non examinés. Par l'hypothèse, nous concluons à l'existence d'un fait tout à fait différent de tout ce qui a pu être observé, à partir de quoi, selon les lois connues, quelque chose d'observé découlerait nécessairement. La première raisonne à partir des particuliers pour aller vers la loi générale ; la seconde va de l'effet à la cause. La première classifie, la seconde explique*[24].

La distinction entre l'induction comme classifieur et l'abduction comme explicateur définit aussi très bien la nature des résultats obtenus par les réseaux neuronaux (et le problème fondamental de l'intelligence artificielle). L'induction statistique complexe effectuée par les réseaux neuronaux se rapproche d'une forme d'abduction faible : bien que de nouvelles catégories et idées se profilent à l'horizon, il apparaît que l'invention et la créativité sont loin d'avoir été automatisées. L'invention de nouvelles règles (définition acceptable de l'intelligence) ne suppose pas seulement la généralisation d'une règle spécifique (comme dans le cas de l'induction et d'abduction faible) mais aussi l'ouverture de plans sémiotiques qui n'étaient au départ pas connectés entre eux ni concevables, comme dans les découvertes scientifiques ou la création de métaphores (abduction forte). Umberto Eco remarquait, dans sa critique de l'intelligence artificielle : « Il n'existe pas d'algorithme pour créer des métaphores, et une métaphore ne peut non plus être produite en fournissant des instructions précises à un ordinateur, quel que soit le volume d'information organisée qu'on y entre[25]. » Eco soulignait que les algorithmes ne sont pas capables de se soustraire au corset des catégories implicitement ou explicitement

incarnées par l'« information organisée » du jeu de données. Inventer des catégories, c'est effectuer un saut, connecter des catégories qui n'avaient jamais été logiquement reliées entre elles. Enfreindre une règle linguistique, c'est inventer une nouvelle règle, mais seulement lorsque cela implique la création d'un ordre plus complexe, dans lequel l'ancienne règle apparaît comme un cas simplifié et primitif. Les réseaux neuronaux peuvent calculer des métaphores *a posteriori*[26], mais ils ne peuvent automatiser *a priori* l'invention de nouvelles métaphores (sans verser dans des résultats comiques, telle la production de texte aléatoire). L'automatisation de l'abduction (forte) demeure la pierre philosophale de l'intelligence artificielle.

L'intelligence artificielle (quasi) explicable

Au fond, l'actuel débat sur l'intelligence artificielle reste dépendant des traumatismes épistémiques suscités par l'essor du calcul neuronal. On prétend que l'intelligence des machines ouvre de nouvelles perspectives de connaissance qui doivent être reconnues comme un patrimoine posthumain (voir la notion lyotardienne d'inhumain), mais on s'intéresse bien peu aux formes symboliques de la reconnaissance de motifs, de l'inférence statistique et de l'abduction faible qui constitueraient un véritable basculement dans le posthumain. On prétend aussi que ces nouvelles échelles de calcul sont une boîte noire échappant au contrôle humain (et politique) sans se rendre compte que l'architecture de cette boîte noire peut être défaite. La suite de ce texte explique que l'humain peut encore s'introduire dans l'abîme « inhumain » du calcul profond et que l'influence humaine demeure reconnaissable dans une bonne part des résultats « inhumains » du calcul.

Il est vrai que les multiples couches superposées de neurones artificiels entrelacent un si grand nombre de calculs qu'il est difficile de discerner dans la structure à quel endroit et de quelle manière une *décision* particulière a été calculée. Les réseaux de neurones artificiels sont considérés comme des « boîtes noires » parce qu'ils ne sont guère, voire pas du tout, capables d'expliquer la causalité ou les caractéristiques importantes pour générer une inférence telle que la classification. Le programmeur n'a souvent aucun contrôle sur les caractéristiques extraites, puisqu'elles sont déduites par le seul réseau neuronal[27].

Une fois de plus, l'armée a clairement perçu le problème. La DARPA (agence de recherche du ministère états-unien de la Défense) a lancé le programme « Explainable Artificial Intelligence » (XAI) pour tenter de résoudre l'effet de boîte noire[28]. Les chercheurs étudient par exemple les scénarios suivants : un char d'assaut autonome qui prend une

direction inattendue, ou la détection imprévue d'armes ennemies dans un paysage neutre. L'idée de la XAI est que les réseaux neuronaux doivent fournir non seulement un output dénué d'ambiguïté mais aussi une justification (relevant du contexte computationnel de cet output). Si, par exemple, la figure d'un ennemi est reconnue (« ceci est l'image d'un soldat armé »), le système dira pourquoi il le pense, c'est-à-dire en fonction de quelles caractéristiques. Des systèmes similaires peuvent s'appliquer à la surveillance des messageries électroniques, pour repérer d'éventuels terroristes, traîtres et agents doubles. Le système tentera non seulement de détecter des anomalies de comportement par contraste avec une configuration sociale normale, mais aussi d'expliquer quel contexte d'éléments permet de décrire une personne comme suspecte. L'automatisation de la détection d'anomalies ayant déjà causé des dégâts (voir l'affaire Skynet au Pakistan[29]), la XAI vise clairement à prévenir de nouvelles catastrophes algorithmiques dans le contexte du *predictive policing*.
L'intelligence artificielle explicable (qu'il serait plus exact d'appeler « *Deep Learning* explicable ») constitue une boucle de contrôle supplémentaire au sommet de l'architecture des réseaux neuronaux et prépare une nouvelle génération de médiateurs épistémiques. Elle implique déjà d'énormes intérêts commerciaux, puisque, par exemple, les compagnies d'assurance ne couvriront que les voitures autonomes disposant d'une « boîte noire computationnelle » qui fournira non seulement des enregistrements audio et vidéo mais aussi la justification de leurs décisions (imaginons le cas du premier accident entre deux véhicules autonomes). Les échelles inhumaines de calcul et l'esthétique du « nouvel âge sombre » ont d'ores et déjà trouvé leurs représentants légaux.

Conclusion

Pour comprendre l'impact historique de l'intelligence artificielle, ce texte a souligné qu'aujourd'hui, le paradigme hégémonique et dominant n'était pas de nature symbolique (la GOFAI) mais connexionniste : les réseaux de neurones qui constituent aussi les systèmes de *Deep Learning*. Ce que les médias dominants appellent « intelligence artificielle » n'est en réalité qu'une désignation folklorique des réseaux neuronaux destinés à la reconnaissance de motifs (une tâche spécifique au sein d'une définition plus large de l'intelligence et qui n'est certainement pas exhaustive). La reconnaissance de motifs est rendue possible par le calcul de l'état interne d'un réseau neuronal qui donne corps à la forme logique de l'induction statistique. L'« intelligence » des réseaux neuronaux n'est par conséquent qu'une inférence statistique de corrélations à partir d'un

jeu de données de formation. Les limites intrinsèques de l'induction statistique résident entre le surapprentissage et l'apophénie, dont les effets se font progressivement jour dans les perceptions collectives et la gouvernance. Quant aux limites extrinsèques de l'induction statistique, on peut les illustrer grâce à la distinction proposée par Peirce, entre induction, déduction et abduction (hypothèse). Si l'induction statistique se rapproche de certaines formes d'abduction faible (par exemple, les diagnostics médicaux), elle est incapable d'automatiser l'abduction forte, telle qu'elle a lieu dans la découverte de lois scientifiques ou dans l'invention de métaphores linguistiques. Tout simplement parce que les réseaux neuronaux ne peuvent franchir la frontière des catégories implicitement intégrées dans le jeu de données de formation. Les réseaux de neurones présentent une autonomie relative dans leurs calculs : ils sont encore dirigés par des facteurs humains et sont les composantes d'un système placé sous le pouvoir des humains. Ils ne montrent absolument aucun signe d'« intelligence autonome » ou de conscience. Les échelles surhumaines de connaissance ne sont acquises qu'avec la collaboration de l'observateur humain, ce qui suggère que l'expression « intelligence augmentée » serait plus précise que celle d'« intelligence artificielle ».
L'inférence statistique *via* les réseaux neuronaux a permis au capitalisme computationnel d'imiter et d'automatiser le travail, peu qualifié comme très qualifié[30]. Personne n'imaginait qu'un conducteur de bus pourrait devenir une source de travail cognitif automatisable par les réseaux de neurones dans les véhicules autonomes.
L'automatisation de l'intelligence reposant sur l'inférence statistique est désormais l'œil que le capital jette sur l'océan de données que constituent le travail, la logique et les marchés globalisés et qui produit des effets nouveaux d'*anormalisation*, cette déformation des perceptions collectives et des représentations sociales que l'on constate dans la manifestation algorithmique des préjugés de classe, de race et de genre[31]. L'inférence statistique est le nouvel œil déformé du Maître du capital[32].

L'auteur souhaite remercier Anil Bawa-Cavia, Nina Franz, et Nikos Patelis pour leurs commentaires.
Traduit de l'anglais par Gauthier Hermann.

Footnotes

1. Frank Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Buffalo, NY, Cornell Aeronautical Laboratory, 1961, p. 28.

2. Umberto Eco, *Semiotics and the Philosophy of Language*. Bloomington, Indiana University Press, 1986, p. 127.

3. Voir Francis Heylighen et Cliff Joslyn, « Cybernetics and Second-Order Cybernetics », *in* R. A.

Meyers (dir.), *Encyclopedia of Physical Science and Technology,* t. 19, New York, Academic Press, 2001.

4. Claude Shannon, « A Mathematical Theory of Communication », *Bell System Technical Journal*, vol. 27, n° 3, 1948.

5. Norbert Wiener, *Cybernetics: Or Control and Communication in the Animal and the Machine*, Cambridge, Mass., MIT Press, 1948, p. 61. Il se trouve que la formulation de Wiener a aussi influencé Jacques Lacan, ainsi que le montre sa conférence de 1955 sur la cybernétique et la psychanalyse, où il interprète les portes logiques littéralement, comme ouvrant ou barrant l'accès à de nouvelles destinées au sein de l'ordre symbolique. Voir Jacques Lacan, « Psychanalyse et cybernétique, ou de la nature du langage », *Le Séminaire, livre II : Le moi dans la théorie de Freud et dans la technique de la psychanalyse*, Paris, Seuil, 1978.

6. Gregory Bateson, *Steps to an Ecology of Mind*, Chicago, University of Chicago Press, 1972.

7. Grégoire Nicolis et Ilya Prigogine, *Self-Organization in Nonequilibrium Systems*, New York, Wiley, 1977.

8. L'intelligence artificielle générale (IAG) tente souvent de trouver un compromis entre l'approche *top-down* (symbolique) et l'approche *bottom-up* (connexionniste), autrement dit de combiner la déduction symbolique avec l'induction statistique. À ce jour, toutefois, seul le paradigme connexionniste des réseaux neuronaux a pu être automatisé avec succès, ce qui jette le doute sur certaines prémisses métaphysiques et centralisantes de l'IAG.

9. Voir Marvin Minsky, *Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain Model Problem*, thèse de doctorat, Princeton University, 1954.

10. Warren McCulloch et Walter Pitts, « A Logical Calculus of the Ideas Immanent in Nervous Activity », *Bulletin of Mathematical Biophysics*, vol. 5, n° 4, 1943 ; et Warren McCulloch et Walter Pitts, « How We Know Universals the Perception of Auditory and Visual Forms », *Bulletin of Mathematical Biophysics*, vol. 9, n° 3, 1947.

11. De façon remarquable, l'ouvrage de Paul Virilio sur la machine de vision s'inspirait aussi du perceptron (mais Virilio ne pouvait pas prévoir que le perceptron deviendrait le paradigme hégémonique de l'intelligence des machines au début du xxie siècle). Voir Paul Virilio, *La Machine de vision : essai sur les nouvelles techniques de représentation*, Paris, Galilée, 1988, chap. 5.

12. Frank Rosenblatt, « The Perceptron, a Perceiving and Recognizing Automaton », Technical Report 85-460-1, 1957, p. 1-2.

13. *Ibid.*, p. 30.

14. Frank Rosenblatt, *Principles of Neurodynamics*, *op. cit.*, p. vii.

15. David Rumelhart et PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2 t., Cambridge, Mass., MIT Press, 1986.

16. Les réseaux de neurones évoluent vers des topologies toujours plus complexes et ont inauguré un véritable art combinatoire (voir les diagrammes de l'auto-encodeur, la machine de Boltzmann, les réseaux de neurones récurrents et à mémoire à court et long termes, les Generative Adversarial Networks, etc.). Les réseaux neuronaux sont les machines les plus articulées et les plus sophistiquées conçues dans la tradition du calcul, dont relèvent l'ancien procédé arabe du

zairia et l'*Ars Magna* (1305) de Raymond Lulle. Voir David Link, « Scrambling T-R-U-T-H: Rotating Letters as a Material Form of Thought », *in* Siegfried Zielinski and Eckhard Fürlus (dir.), *Variantology 4. On Deep Time Relations of Arts, Sciences and Technologies in the Arabic-Islamic World*, Cologne, König, 2010.

17. Frank Rosenblatt, *Principles of Neurodynamics*, *op. cit.*, p. 9.

18. Cf. Ethem Alpaydin, *Introduction to Machine Learning*, 2e éd., Cambridge, Mass., MIT Press, 2014, p. 260.

19. Ce n'est là qu'un exemple spécifique, et donné afin d'illustrer mon propos. Les fonctions d'activation opèrent aussi de manières différentes.

20. Concernant les tentatives d'automatiser l'abduction faible, voir « Automatic Abductive Scientists », *in* Lorenzo Magnani, *Abductive Cognition*, Springer Science & Business Media, 2009, p. 112.

21. Charles S. Peirce, « Some Consequences of Four Incapacities » (1868), *in* Nathan Houser et Christian Kloesel (éd.), *The Essential Peirce*, t. I : *1867-1893*, Bloomington, Indiana University Press, 1992, p. 54.

22. Le mot « ontologie » est ici utilisé dans le sens où l'entend la science de l'information.

23. Charles S. Peirce, *Collected Papers*, Cambridge, Mass., Belknap Press, 1965, t. V, p. 145.

24. Charles S. Peirce, « Deduction, Induction, and Hypothesis » (1878), *in* Nathan Houser et Christian Kloesel (éd.), *The Essential Peirce*, t. I, *op. cit.*, p. 194.

25. Umberto Eco, *Semiotics and the Philosophy of Language*, *op. cit.*, p. 127.

26. Voir Word2Vec, application destinée à cartographier l'intégration de mots dans l'espace vectoriel.

27. Cela est vrai aussi bien pour l'apprentissage dirigé que pour l'apprentissage non dirigé. Je remercie Anil Bawa-Cavia de m'avoir précisé ce point.

28. Voir www.darpa.mil/program/explainable-artificial-intelligence

29. Matteo Pasquinelli, « Arcana Mathematica Imperii: The Evolution of Western Computational Norms », *in* Maria Hlavajova *et al.* (dir.), *Former West*, Cambridge, Mass., MIT Press, 2017.

30. Il existe différentes approches de l'intelligence des machines, mais le connexionnisme jouit d'une hégémonie patente dans l'automatisation. Pour une introduction accessible à l'apprentissage automatique, voir Pedro Domingos, *The Master Algorithm*, New York, Basic Books, 2015.

31. Matteo Pasquinelli, « Abnormal Encephalization in the Age of Machine Learning », *e-flux*, n° 75, septembre 2016.

32. « La dextérité et la minutie du travailleur sur machine vidé de sa substance en tant qu'individu disparaissent tel un minuscule accessoire devant la science, devant les énormes forces naturelles et le travail de masse, dont le système des machines est l'incarnation et qui fondent avec lui la puissance du "maître" (*master*). » Karl Marx, *Le Capital*, livre I, trad. collective dirigée par J.-P. Lefebvre, Paris, Presses universitaires de France, 1993, p. 475.

Matteo Pasquinelli is Professeur en Théorie des Medias à l'Université d'Art et de Design, Karlsruhe.