HAL
open science

La revanche des neurones : L'invention des machines inductives et la controverse de l'intelligence artificielle

Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières

▶ **To cite this version:**

Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières. La revanche des neurones : L'invention des machines inductives et la controverse de l'intelligence artificielle : L'invention des machines inductives et la controverse de l'intelligence artificielle. Réseaux : communication, technologie, société, 2018, 5 (211), pp.173-220. 10.3917/res.211.0173 . hal-02005537

HAL Id: hal-02005537

https://sciencespo.hal.science/hal-02005537

Submitted on 4 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

LA REVANCHE DES NEURONES

L'invention des machines inductives et la controverse de l'intelligence artificielle

Dominique CARDON
Jean-Philippe COINTET
Antoine MAZIÈRES

La Découverte | « Réseaux »
2018/5 n° 211 | pages 173 à 220
ISSN 0751-7971
ISBN 9782348040689

Pour citer cet article :
Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones. L'invention des machines inductives et la controverse de l'intelligence artificielle », Réseaux 2018/5 (n° 211), p. 173-220.
DOI 10.3917/res.211.0173

Résumé
Depuis 2010, les techniques prédictives basées sur l'apprentissage artificiel (machine learning), et plus spécifiquement des réseaux de neurones (deep learning), réalisent des prouesses spectaculaires dans les domaines de la reconnaissance d'image ou de la traduction automatique, sous l'égide du terme d'"Intelligence artificielle". Or l'appartenance de ces techniques à ce domaine de recherche n'a pas toujours été de soi. Dans l'histoire tumultueuse de l'IA, les techniques d'apprentissage utilisant des réseaux de neurones - que l'on qualifie de "connexionnistes" - ont même longtemps été moquées et ostracisées par le courant dit "symbolique". Cet article propose de retracer l'histoire de l'Intelligence artificielle au prisme de la tension entre ces deux approches, symbolique et connexionniste. Dans une perspective d'histoire sociale des sciences et des techniques, il s'attache à mettre en évidence la manière dont les chercheurs, s'appuyant sur l'arrivée de données massives et la démultiplication des capacités de calcul, ont entrepris de reformuler le projet de l'IA symbolique en renouant avec l'esprit des machines adaptatives et inductives de l'époque de la cybernétique.

Mots-clés
Réseaux de neurones, Intelligence artificielle, Connexionnisme, Système expert, Deep learning

---

Abstract
Since 2010, machine learning based predictive techniques, and more specifically deep learning neural networks, have achieved spectacular performances in the fields of image recognition or automatic translation, under the umbrella term of "Artificial Intelligence". But their filiation to this field of research is not straightforward. In the tumultuous history of AI, learning techniques using so-called "connectionist" neural networks have long been mocked and ostracized by the "symbolic" movement. This article retraces the history of artificial intelligence through the lens of the tension between symbolic and connectionist approaches. From a social history of science and technology perspective, it seeks to highlight how researchers, relying on the availability of massive data and the multiplication of computing power have undertaken to reformulate the symbolic AI project by reviving the spirit of adaptive and inductive machines dating back from the era of cybernetics.

Keywords
Neural networks, Artificial intelligence, Connectionism, Expert systems, Deep learning

L'épisode est en passe de devenir légendaire dans l'histoire de l'informatique. Nous sommes en octobre 2012 à la conférence ECCV qui réunit les chercheurs spécialisés en computer vision[1].

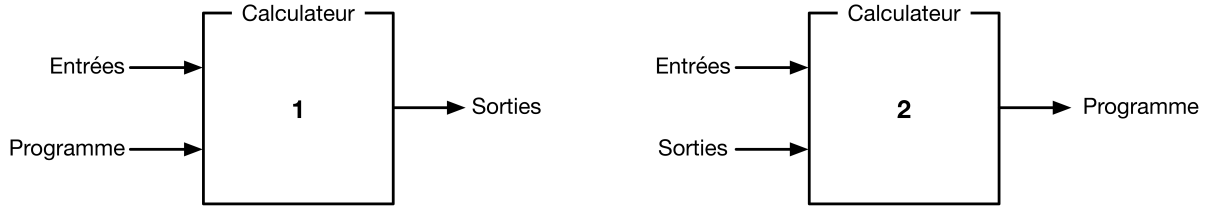
> « Alors à la compétition de 2012, qui débarque ? C'est Hinton [*le "père" du renouveau des réseaux de neurones*] et c'est le séisme. Il ne connaît rien au domaine de la vision par ordinateur et il prend deux petits gars pour tout faire sauter ! Un [*Alex Krizhevsky*] qu'il a enfermé dans une boîte et il lui a dit : "Tu ne sors pas tant que ça ne marche pas !" Il a fait tourner des machines énormes, qui avaient des GPU à l'époque qui n'étaient pas ultra, mais qu'il faisait communiquer entre eux pour les booster. C'était un truc de machinerie complètement dingue. Sinon, ça n'aurait jamais marché, un savoir-faire de geek, de programmation qui est hallucinant. À l'époque, les mecs de computer vision s'excitaient sur ImageNet depuis deux trois ans [*une base de données de 1,2 million d'images étiquetées avec 1 000 catégories servant de benchmark pour comparer les résultats en classification des différents compétiteurs*]. Le number one, il était à 27,03 % d'erreur, le number 2 à 27,18 %, le number 3 à 27,68 %. Et Hinton, il envoie son mec sorti de nulle part : "on a fait tourner un gros deep, on est à 17 !". Il met 10 points à tout le monde ! Comme ça, le jeune geek, il arrive, il annonce le résultat, la salle bondée à craquer. Enfin, il comprend rien à rien, genre il a 17 ans ! Il ne sait pas pourquoi les trucs sont là. Lui, il était enfermé dans sa boîte, il ne connaissait rien au domaine. Et là, il est face à Fei Fei ! Et tu as LeCun qui est assis au fond de la salle qui se lève pour répondre aux questions [*Li Fei Fei, professeur d'informatique qui dirige SAIL le laboratoire historique d'intelligence artificielle de Stanford ; Yann LeCun, aujourd'hui directeur de FAIR, le laboratoire d'intelligence artificielle de Facebook et un des acteurs centraux du renouveau des réseaux de neurones*]. Et tu as tous les grands manitous du computer vision qui essayent de réagir : "Mais en fait c'est pas possible. Ça va pas marcher pour la reconnaissance d'objet quand il faut..." Enfin, les mecs étaient tous par terre parce que grosso modo cela foutait en l'air 10 ans d'intelligence, de tuning, de sophistication.

[1] Cette enquête a été conduite dans le cadre du projet ALGODIV (ANR-15-CE38-0001). Les auteurs souhaitent remercier Telmo Menezes pour ses conseils. Dans cette enquête, nous exploitons trois entretiens réalisés avec des chercheurs français en informatique qui ont participé à la renaissance des réseaux de neurones. Afin de conserver le caractère brut de leurs propos, ils ont été anonymisés.

> C'est pas forcément des gens qui font de la logique formelle, mais c'est des gens qui sont quand même dans cette idée qu'il faut comprendre, qu'il faut savoir expliquer pourquoi on met les branches comme ça et qu'on raisonne comme ça, et que l'on avance comme ça, et qu'il faut toute cette intelligence des features qui va avec et qui aide à dire que l'on comprend parfaitement ce que l'on fait et que l'on sait pourquoi c'est là. Et le mec il arrive avec une grosse boîte noire de deep, il a 100 millions de paramètres dedans, il a entraîné ça et il explose tout le domaine. "Est-ce que vos modèles sont invariants si l'image bouge ?" Le gars, il a même pas compris la question ! C'est LeCun qui répond : "Alors, ces modèles sont invariants parce que..." Il était trop content, parce que Fei Fei lui demande : "Mais Yann, est-ce que ces modèles sont fondamentalement différents des modèles que tu as inventé dans les années 1980 ?" Et là Yann, il peut dire : "Nan, c'est exactement les mêmes et on a gagné toutes les compétitions avec !"[2]. »

Ce récit coloré de l'annonce des performances en classification d'images d'une technique de *deep learning* (Krizhevsky, Sutskever et Hinton, 2012) témoigne des effets que provoque sur une communauté scientifique la réussite soudaine d'un paradigme hétérodoxe longtemps marginalisé[3]. Surprise devant le résultat, interrogation sur la validité épistémique de la nouvelle démarche, inquiétude sur le devenir du paradigme orthodoxe, moquerie devant l'ignorance des enjeux théoriques du domaine des nouveaux entrants, vertige face au renversement de paradigme qui se profile... Depuis 2010, domaine après domaine, les réseaux de neurones profonds provoquent la même perturbation au sein des communautés informatiques traitant du signal, de la voix, de la parole ou du texte. Une méthode d'apprentissage proposant le traitement le plus « brut » possible des entrées, évacuant toute modélisation explicite des caractéristiques des données et optimisant la prédiction à partir d'énormes échantillons d'exemples, produit de spectaculaires résultats. Une manière simple de figurer ce renversement est de le caractériser comme le passage d'une machine hypothético-déductive à une machine inductive (figure 1).

Figure 1. Machine hypothético-déductive (1) et machine inductive (2)



Ce qui était conçu comme la partie « humaine » de la fabrication des calculateurs, le programme, les règles ou le modèle, n'est plus ce qui est introduit dans le système, mais ce qui en résulte. Le regard que portent les sciences sociales sur ce tournant inductif consiste souvent à déconstruire l'illusion naturaliste des données « brutes » et les naïvetés d'un calcul sans théorie (Gitelman, 2013). Si une telle mise en garde est certainement nécessaire pour

[2] Interview V., chercheur en vision par ordinateur, 12 mars 2018.
[3] Y. LeCun a livré sa version du même événement dans une vidéo (à partir de 20') : « Heroes of Deep Learning: Andrew Ng interviews Yann LeCun », YouTube, 7 avril 2018.

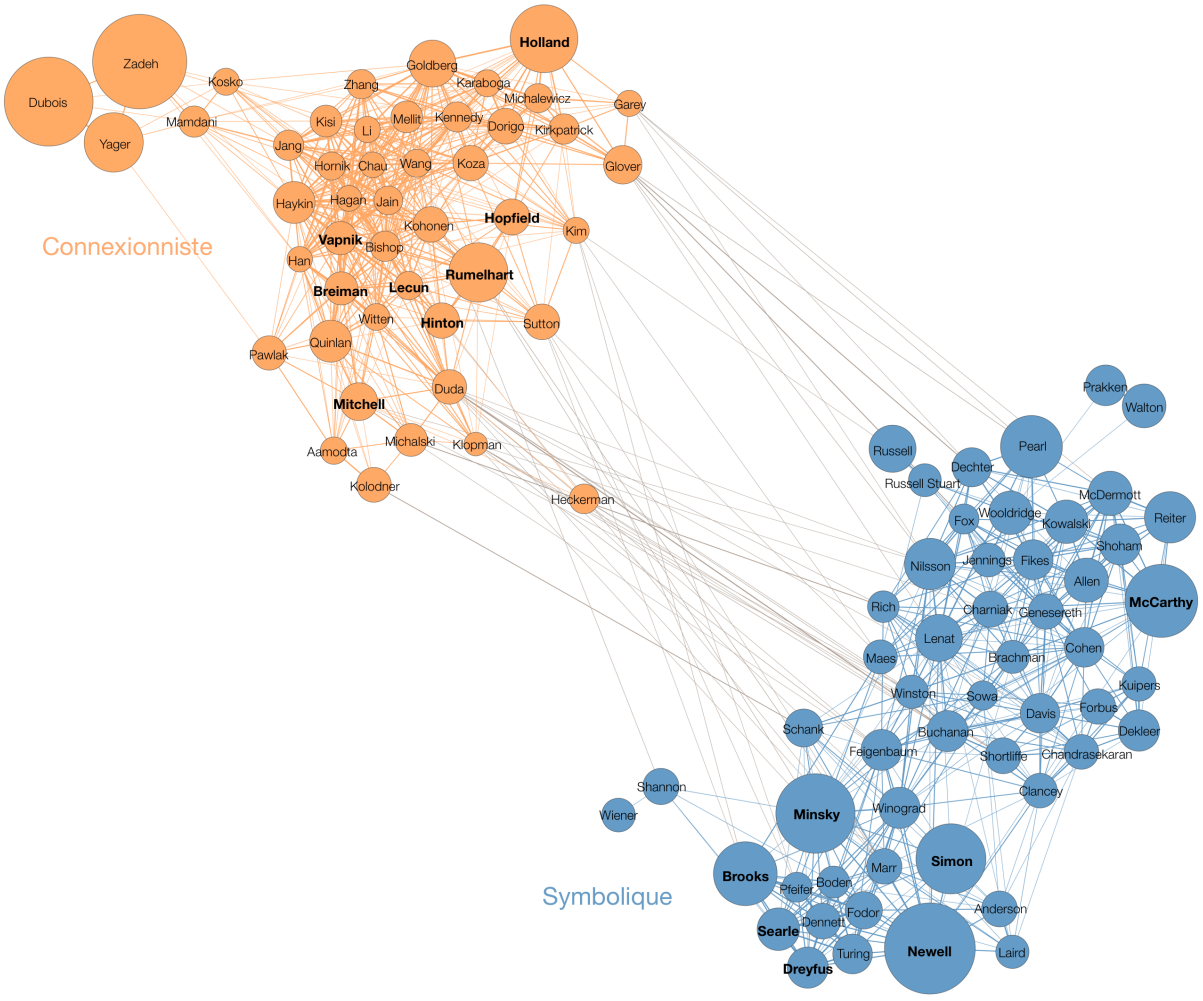
relativiser certains discours imprudents assurant que les « données parlent d'elles-mêmes », elle ne rend cependant pas justice au travail résolu et intensément artificiel entrepris par les promoteurs des techniques de *deep learning* pour imposer la seconde architecture de calcul, celle que nous appellerons dans cet article *machine inductive* et, plus précisément encore, *machine connexionniste* afin de mettre en évidence le type particulier d'induction dont elle se réclame. La fabrication d'artefacts susceptibles de produire un calcul inductif sur de grandes masses de données est le résultat d'une histoire conflictuelle et d'une série de constructions d'une très grande ingéniosité. L'induction est quelque chose qu'il faut constamment amener à la machine, défendre contre des opposants, produire au moyen de calculs spécifiques, déployer dans des architectures propres, calibrer avec des données adaptées. Les concepteurs de ces machines ne sont pas les naturalistes candides auxquels les sciences sociales constructivistes aiment parfois les réduire. L'idée de confier aux machines le soin de produire des prédictions pertinentes en apprenant des données – *i.e.* le calcul inductif – est un projet, une théorie et, surtout, un dispositif qui a une histoire agitée. Pour être mis en place et produire ses effets, il a exigé un patient travail de reconfiguration de l'architecture des machines « intelligentes » dont cet article voudrait rendre compte.

Symbolique vs Connexionnisme

La méthode des réseaux de neurones que l'on vient de voir triompher à ECCV'12 n'est en rien nouvelle. Profitant de l'augmentation de la puissance de calcul des ordinateurs et de l'accessibilité de gigantesques bases de données, elle parvient aujourd'hui à honorer une promesse qui avait été faite au début de la cybernétique. De façon surprenante, le terme récemment retenu pour qualifier ces stupéfiantes prouesses calculatoires est celui d'*intelligence artificielle* (IA). Le retour sur le devant de la scène de ce vocable forgé en 1956 par John McCarthy constitue une intéressante énigme pour l'histoire des sciences et des techniques. La plupart des observateurs rigoureux soulignent en effet que c'est dans le seul domaine des méthodes d'apprentissage et, notamment, de l'apprentissage profond (*deep learning*), que des progrès sensibles de la prédiction calculée ont lieu actuellement. Or l'appartenance de ces techniques au champ de l'IA n'est pas toujours allée de soi. Dans l'histoire tumultueuse de ce domaine de recherche, les techniques d'apprentissage utilisant des réseaux de neurones – que l'on appellera donc « connexionnistes » – ont même longtemps été moquées et ostracisées par le courant « symbolique ». La tension entre ces deux approches est née avec la démarcation que la naissance de l'Intelligence artificielle opposa à la première cybernétique. L'approche *symbolique* qui constitue le cadre de référence initial de l'IA s'est identifiée à un cognitivisme orthodoxe : penser, c'est calculer des symboles qui ont à la fois une réalité matérielle et une valeur sémantique de représentation. En revanche, le paradigme connexionniste considère que penser s'apparente à un calcul massivement parallèle de fonctions élémentaires – celles qui seront distribuées au sein d'un réseau de neurones – dont les comportements signifiants n'apparaissent au niveau collectif que comme un effet émergent des interactions produites par ces opérations élémentaires (Andler, 1992). Cette distinction entre deux manières de concevoir et de programmer le fonctionnement « intelligent » d'une machine est au principe d'une tension qui n'a jamais cessé de structurer très profondément les orientations de recherche, les trajectoires scientifiques et la conception d'infrastructure de calcul. Aussi assiste-t-on aujourd'hui à un de ces retournements de situation dont l'histoire des sciences et des techniques est

coutumière : une stratégie de recherche marginalisée par ceux qui ont contribué à poser les cadres conceptuels de l'Intelligence artificielle revient au-devant de la scène et se trouve désormais en position de redéfinir très différemment le domaine dont elle avait été exclue. Comme le souligne ironiquement Michael Jordan (2018), « c'est l'agenda intellectuel de Wiener qui domine aujourd'hui sous la bannière de la terminologie de McCarthy ».

Figure 2. Réseau de co-citations des 100 auteurs les plus cités par les publications scientifiques mentionnant « Artificial Intelligence »[4]



Pour faire le récit de ce chassé-croisé, il est d'abord possible d'en dessiner la chronologie à partir des publications scientifiques rassemblées dans le *Web of Science* (WoS). Il suffit en effet d'observer le réseau de co-citations des auteurs les plus cités parmi les articles mentionnant « Artificial Intelligence » pour qu'apparaisse le clivage entre chercheurs investis dans les approches symbolique ou connexionniste. On peut ainsi voir sur la figure 2 les noms des principaux acteurs que nous rencontrerons dans cet article se distribuer clairement selon leur communauté. Au centre des « connexionnistes », Rumelhart, LeCun et Hinton figurent le noyau fondateur du *deep learning* et sont accompagnés des chercheurs qui, à des époques

[4] Le corpus « Intelligence Artificielle » contient 27 656 publications rassemblées en février 2018 sur *Web of Science* par la requête : TS=("artificial intelligence"). La taille des nœuds dépend de la fréquence avec laquelle l'auteur apparaît. Les auteurs régulièrement cocités dans les mêmes publications sont liés dans le réseau. Un algorithme de détection de communauté révèle la bipartition du réseau en deux communautés cohésives.

différentes (Holland, Hopfield), ont nourri ce mouvement ainsi que les principaux contributeurs aux multiples méthodes du *machine learning*, comme Breiman, Mitchell ou Vapnik. Du côté « symbolique », on retrouve le noyau fondateur de l'IA (McCarthy, Minsky, Simon et Newell) dans une disposition qui reflète leurs proximités et leurs divergences, entouré par les principaux contributeurs à la production des modélisations cognitives, des systèmes experts et même à la critique de l'IA symbolique (Dreyfus, Searle, Brooks).

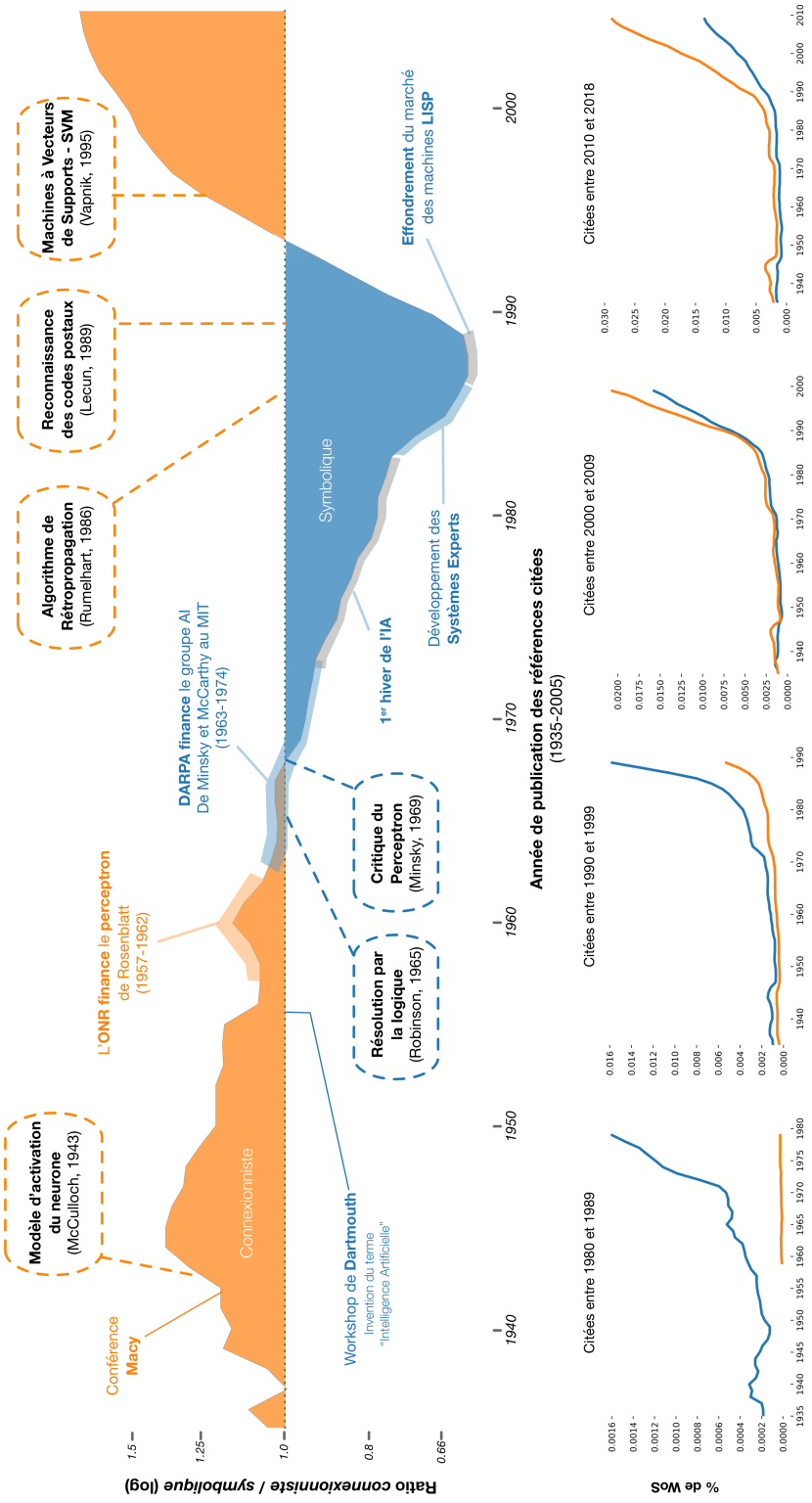
Mais la controverse entre les deux communautés de l'IA apparaît mieux si l'on observe la chronologie de l'impact académique des publications scientifiques des courants symbolique[5] et connexionniste[6] de 1935 à 2005. La frise chronologique ci-dessous (figure 3) fait apparaître la naissance du paradigme connexionniste avec la première cybernétique. Puis, à partir du début des années 1960, c'est le paradigme symbolique qui va dominer et définir les principales caractéristiques de l'IA. Ce n'est enfin qu'au milieu des années 1990, après le deuxième hiver de l'IA, que le paradigme connexionniste redevient nettement dominant dans les publications scientifiques sous la bannière du *deep learning*.

Pour retracer cette histoire, nous proposons une grille de lecture très simple afin d'isoler, au sein d'un vaste ensemble de technologies hétérogènes et d'un très haut niveau de complexité, un certain nombre de points de repère permettant de rendre compte simultanément de la transformation des infrastructures de calcul et des différentes manières de problématiser leur performance. Pour observer ensemble la conception des systèmes techniques et leur visée épistémique, on considérera qu'une machine « intelligente » doit articuler selon des configurations différentes un *monde*, un *calculateur* et un *horizon*. Ces notions se réfèrent au cadre fonctionnel dans lequel est habituellement décomposé le design des artefacts intelligents selon des terminologies variées : « environnement »/« entrées »/« données »/«base de connaissances » (*monde*), « calcul »/« programme »/« modèle »/« agent » (*calculateur*) et « objectifs »/« résultats »/« sorties » (*horizon*). On dira donc que les machines prédictives installent un *calculateur* sur un *monde* en lui conférant un *horizon*. Les dispositifs conçus tout au long de l'histoire de l'IA équipent le *monde*, le *calculateur* et l'*horizon* d'entités variées et changeantes. Ils proposent aussi des manières radicalement différentes d'articuler l'architecture de ces ensembles. Le déplacement des recherches en IA d'une *machine symbolique* vers une *machine connexionniste* n'est donc pas la conséquence d'une mutation de l'histoire des idées, ou de la validité d'un modèle scientifique sur un autre, mais le résultat d'une controverse qui a conduit les acteurs à déplacer, transformer et redéfinir profondément la forme donnée à leurs artefacts. Le processus auquel cette grille de lecture nous permet d'être attentifs est un long travail historique de recomposition des alliances et des paradigmes entre communautés scientifiques en compétition. Celui-ci affecte les techniques de calcul, mais aussi et surtout la forme donnée à ces machines, leurs objectifs, les données dont elles traitent et les questions qu'elles adressent (Latour, 1987). Pour le dire

[5] Le corpus « Symbolique » contient 65 522 publications rassemblées en février 2018 sur Web of Science par la requête: TS=("knowledge representation*" OR "expert system*" OR "knowledge based system*" OR "inference engine*" OR "search tree*" OR "minimax" OR "tree search" OR "Logic programming" OR "theorem prover*" OR ("planning" AND "logic") OR "logic programming" OR "lisp" OR "prolog" OR "deductive database*" OR "nonmonotonic reasoning*").

[6] Le corpus « Connexionniste » contient 106 278 publications rassemblées en février 2018 sur Web of Science par la requête: TS=("artificial neural network*" OR "Deep learning" OR "perceptron*" OR "Backprop*" OR "Deep neural network*" OR "Convolutional neural network*" OR ("CNN" AND "neural network*") OR ("LSTM" AND "neural network*") OR ("recurrent neural network*" OR ("RNN*" AND "neural network*")) OR "Boltzmann machine*" OR "hopfield network*" OR "Autoencoder*" OR "Deep belief network*" OR "recurrent neural network*").

Figure 3. Évolution de l'influence académique des approches connexionniste et symbolique



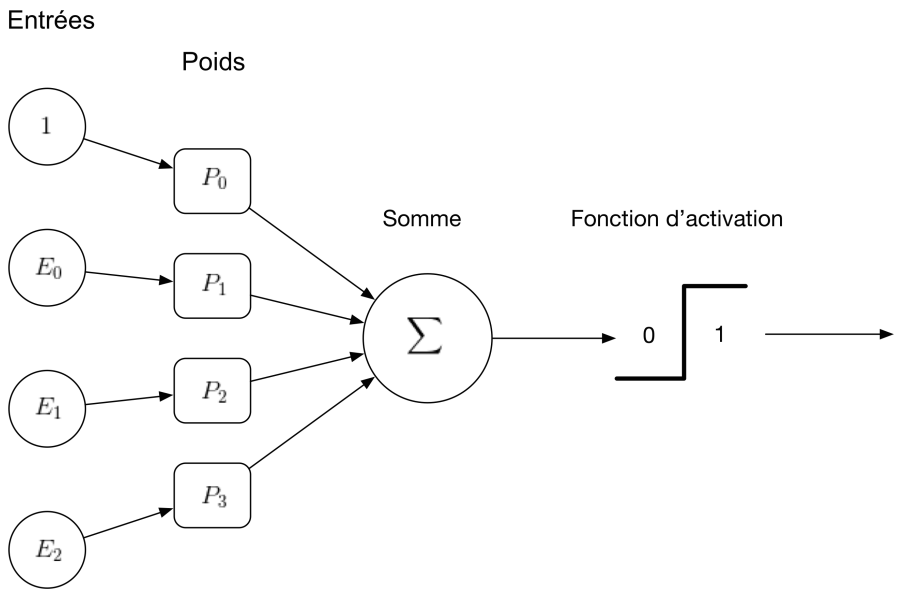
La courbe principale (en haut) représente l'évolution du ratio entre le nombre de publications citées dans le corpus connexionniste (en clair) et le nombre correspondant dans le corpus symbolique (en foncé), tous deux normalisés par le nombre total de publications dans WoS. Les courbes annexes (en bas) représentent pour chacun des corpus, le nombre de publications citées au cours d'une période donnée.

dans une formule que l'on précisera tout au long de l'article : alors que les concepteurs des machines symboliques cherchaient à insérer dans le *calculateur* et le *monde* et l'*horizon*, la réussite actuelle des machines connexionnistes tient au fait que de façon presque opposée, ceux qui les fabriquent vident le *calculateur* pour que le *monde* se donne à lui-même son propre *horizon*.

LA CYBERNÉTIQUE ET LE PREMIER CONNEXIONNISME

Les réseaux de neurones trouvent leur origine dans l'histoire pionnière de l'informatique et de la première cybernétique. Bien que l'étiquette soit postérieure, celle-ci peut en effet être dite « connexionniste »[7] et ne cessera de se référer à la proposition de modéliser mathématiquement un réseau de neurones faite par le neurophysiologiste Warren McCulloch et le logicien Walter Pitts en 1943. Cet article fondateur continue, jusque dans les citations actuelles des articles de *deep learning*, à être donné comme le point de départ de l'aventure connexionniste. La frise chronologique de l'activité scientifique en IA (figure 3) fait clairement apparaître la domination de l'approche connexionniste pendant la période de la première cybernétique. L'article *princeps* de McCulloch et Pitts propose un modèle formel (figure 4) dans lequel le neurone prend des variables en entrées, y applique un poids pour produire une somme qui, si elle dépasse un certain seuil, déclenche l'activation du neurone.

Figure 4. Modèle formel d'un neurone artificiel à seuil binaire



Cette proposition n'est pas formulée comme relevant de l'intelligence artificielle – le terme n'existe pas –, mais comme un outil d'expérimentation en neurophysiologie fidèle aux connaissances biologiques de l'époque sur les processus neuronaux du cerveau. Elle va être rapidement associée à l'idée d'apprentissage à travers les travaux du neuropsychologue Donald O. Hebb (1949) qui montrent que l'activation répétée d'un neurone par un autre, à

[7] Le premier emploi du terme de « connexionnisme » apparaît chez D. Hebb en 1949 et sera ensuite repris par F. Rosenblatt en 1958 (Andler, 1992).

travers une synapse donnée, augmente sa conductivité et peut être considérée comme un apprentissage. Bio-inspiré, le modèle du neurone formel constitue alors un des principaux points de réflexion de la cybernétique et va devenir la pièce centrale du calculateur des premières machines « intelligentes » (Dupuy, 2005).

Un couplage étroit entre le monde et le calculateur

Ce qui caractérise l'architecture de ces machines est que leur couplage avec l'environnement (*le monde*) est si intime qu'il n'est pas nécessaire de doter leur calculateur d'une agentivité propre. La proposition de la cybernétique est d'en faire une simple boîte noire apprenante et associationniste dont l'horizon se règle en mesurant l'écart (*i.e.* l'erreur) entre le monde et le comportement de la machine. Cette représentation de la machine intelligente s'appuie d'abord sur une conception matérialiste de l'information qui se distingue de celle, symbolique, qui va prévaloir au moment de la naissance de l'Intelligence artificielle (Triclot, 2008). Quantité d'ordre opposée à l'entropie, l'information est un signal avant d'être un code. Avec la théorie de l'information développée par Shannon (1948), celle-ci n'a pas besoin d'être associée à une quelconque signification. L'information y est conçue comme une forme pure, indépendamment de toute autre considération, elle se limite à « exprimer la quantité d'ordre ou de structure dans un agencement matériel » (Triclot, 2008).

La machine cybernétique ne définit ensuite l'*horizon* de son calcul que d'une comparaison des entrées et des sorties vers le *monde*. Le dispositif prédictif appliqué au guidage des missiles anti-aériens de Norbert Wiener (1948) repose sur la mise à jour continue de leur trajectoire en comparant le trajet effectif de la cible avec les précédentes estimations. L'appareil doit converger vers la meilleure solution en fonction des données disponibles ; celles-ci nourrissent, corrigent et orientent le calculateur. Le *feed-back négatif* – *i.e.* l'incorporation de la mesure de l'erreur en sortie comme une nouvelle entrée d'un système adaptatif – va ainsi constituer le principal axiome de la cybernétique. C'est lui qui permet de considérer les systèmes techniques sous une forme strictement comportementaliste, en écho à la psychologie behavioriste de l'époque (Skinner, 1971). En continuité avec les organismes vivants, la machine s'adapte inductivement aux signaux de l'environnement dans un couplage si étroit qu'il ne demande pas de lui donner, en interne, des représentations ou des intentions, bref une « intelligence » qui lui soit propre. Lorsqu'Arturo Rosenblueth, Norbert Wiener et Julian Bigelow (1943) formulent les principes de base de la cybernétique, ils imaginent une machine autocorrective capable, grâce à des opérateurs probabilistes, de modifier ou de se donner des finalités qui ne sont pas « internes », mais qui sont produites par l'adaptation de son comportement au vu des erreurs qu'elle commet. De façon rigoureusement « éliminativiste », la conception des machines cybernétiques peut faire l'économie des notions d'intention, de plan ou de raisonnement (Galison, 1994). En théorisant le fonctionnement de l'une des plus notoires d'entre elles, l'Homeostat, Ross Ashby (1956, p. 110) décrira comme une « boîte noire » la partie calculatrice du système environnement/machine[8]. La configuration des machines prédictives de la cybernétique couple si étroitement le *monde* et le *calculateur* que son *horizon* est une optimisation du fonctionnement adaptatif du système qu'ils forment ensemble. Les machines cybernétiques des années 1950 (Homeostat, Adaline, etc.) ne seront que des artefacts de laboratoire à

[8] Sur l'Homeostat, voir Pickering (2010) et Rid (2016).

ambition et à capacité de calcul très réduites, en revanche les calculateurs du *deep learning* parviendront de façon beaucoup plus efficace à poser une boîte noire sur un monde de données en faisant des sorties des entrées.

Le perceptron et les machines connexionnistes

Les réseaux de neurones de McCulloch et Pitts fournissent, notamment dans le domaine de la reconnaissance visuelle, une solution particulièrement adaptée pour équiper le calculateur de ces premières machines adaptatives. À la fin des années 1950, elles connaissent un développement important qui va participer à la première vague d'intérêt public pour les machines-cerveaux[9]. L'approche connexionniste inspire les travaux de Bernard Widrow (Adaline), de Charles Rosen à Stanford (Shakey) ou même Pandemonium, le dispositif hybride d'Oliver Selfridge (1960). C'est cependant l'initiative du Perceptron de Frank Rosenblatt (1957-1961), psychologue et informaticien à l'université de Cornell, qui incarne la première véritable machine connexionniste et deviendra l'emblème d'une autre manière de donner un comportement intelligent à un artefact calculatoire. Ce dispositif conçu pour la reconnaissance d'image reçoit beaucoup d'attention et obtient un financement important de la marine américaine (ONR). La machine imaginée par Frank Rosenblatt s'inspire des réseaux de neurones formels de McCulloch et Pitts, tout en y ajoutant un mécanisme d'apprentissage. Dans les couches superposées du Perceptron, les neurones d'entrée simulent l'activité de la rétine et les neurones de sorties classent les « traits » reconnus par le système ; seules les couches cachées, intermédiaires, sont capables d'apprentissage. À la différence de l'organisation logique – et « descendante » – de McCulloch et Pitts, Frank Rosenblatt revendique une démarche « bottom-up » qui laisse le mécanisme d'apprentissage organiser de façon statistique la structure du réseau. Après une première implémentation logicielle, Frank Rosenblatt entamera la construction de l'unique version matérielle du Perceptron : le Mark I, qui regroupe 400 cellules photoélectriques connectées à des neurones. Les poids synaptiques étaient encodés dans des potentiomètres, et les changements de poids pendant l'apprentissage étaient effectués par des moteurs électriques. La mise en œuvre concrète de ces machines apprenantes restera cependant très rare en raison des limitations techniques de l'époque et, surtout, elle se verra stoppée par le développement d'une IA explorant une direction de recherche tout autre, « symbolique ».

L'IA SYMBOLIQUE

Lorsque les principaux promoteurs de la réunion fondatrice de Dartmouth, John McCarthy et Marvin Minsky, lancent en 1956 le terme d'« intelligence artificielle » (IA), c'est pour l'opposer au connexionnisme de la première cybernétique (Dupuy, 2005)[10]. Il s'agit très explicitement de donner aux machines une autre ambition que celle d'un ajustement adaptatif des entrées et des sorties. L'ambition de l'IA « symbolique »[11] est de mettre dans les ordinateurs, à travers leurs programmes, des règles permettant de manipuler des

[9] Il faut souligner qu'au début des années 1960, les travaux sur les réseaux de neurones sont alors considérés comme une voie potentielle de l'IA. Ils seront très vite minoritaires avant d'être complètement marginalisés au sein du domaine naissant, mais les grandes conférences du début des années 1960 réunissent encore des chercheurs du courant symbolique et du courant connexionniste (Anderson et Rosenfeld, 1988).

[10] Sur l'histoire de la première époque de l'IA, voir Crevier (1997), McCorduck (1979) et Nilsson (2010).

[11] Aussi appelée LGAI pour *Logic-Based AI*, AGI (*artificial general intelligence*), « IA forte » (*strong* ou *full AI*) et aujourd'hui « IA à l'ancienne » (*Good old-fashioned AI*) (Haugeland, 1985).

représentations de haut niveau. La naissance de l'IA est ainsi apparue comme un véritable front « anti-inductif » dans lequel la logique devait contrer les « chimères » de l'approche connexionniste accusée de s'être refusée à définir un traitement de l'information indépendant des processus physiques et de proposer une théorie de l'esprit (Minsky, 1986)[12]. Comme le montre la frise chronologique (figure 3), c'est l'approche symbolique qui va dominer la production scientifique en IA du milieu des années 1960 aux débuts des années 1990.

Celle-ci a d'abord été nourrie par les travaux d'Herbert Simon conduits avec Allen Newell à la Rand dans les années 1950. En 1956, ils écrivent le premier programme destiné à simuler la prise de décision par une machine, le Logic Theorist (1956) en annonçant – ce qui deviendra une habitude caractéristique des chercheurs en IA – que « d'ici Noël, Allen Newell et moi aurons inventé une machine pensante » (McCorduck, 1979, p. 116). La modélisation du raisonnement est la caractéristique centrale de cette première vague de l'IA qui s'étend de 1956 au début des années 1970. Le domaine de recherche est rapidement constitué par un groupe restreint réunissant le MIT (Minsky, Papert), Carnegie Mellon (Simon, Newell) et Stanford University (McCarthy). En dépit de divergences internes, ce cercle fermé s'arroge le monopole de la définition des enjeux de l'IA, capture l'essentiel de – considérables – financements et l'accès aux grands systèmes informatiques. De 1964 à 1974, ils vont recevoir 75 % du financement des recherches en IA distribué par l'ARPA et l'Air Force (Fleck, 1982, p. 181) et bénéficier des rares capacités de calcul nécessaires pour leurs projets. À l'ARPA, ils bénéficient du soutien sans faille de Joseph Licklider qui finance les projets symboliques tout en les justifiant par d'hypothétiques applications militaires.

Cette prise de pouvoir du courant symbolique sur la définition alors floue et très ouverte des machines intelligentes prendra la forme d'une excommunication prononcée dans le livre que Marvin Minsky et Samuel Papert (1969) consacrent à démontrer l'inefficacité des réseaux de neurones. Au début des années 1960, les approches connexionnistes héritières de la première cybernétique connaissaient un certain engouement porté par le succès médiatique du Perceptron de Frank Rosenblatt. Bien que, étudiant, il ait lui-même développé des réseaux de neurones (Snarc, 1951), Marvin Minsky souhaite affirmer la prééminence mathématique de l'IA symbolique face au caractère « mystique », « entouré d'une atmosphère romantique » des systèmes autoorganisés et distribués des connexionnistes (Minsky et Papert, 1969, note 13). En prenant pour cible une version réduite et simplifiée du Perceptron à une couche, il démontre avec Simon Papert que les réseaux de neurones sont incapables de calculer la fonction XOR (le OU exclusif) et qu'ils sont donc sans avenir. Comme l'a montré Mikel Olazaran (1996), la stratégie de Minsky et Papert est de consacrer la prééminence du courant symbolique dans la définition de l'Intelligence artificielle. Même si les effets du livre ont sans doute dépassé les intentions de leurs auteurs, sa conséquence sera sans appel : après la mort prématurée de Frank Rosenblatt en 1971, les réseaux de neurones seront abandonnés, leur financement arrêté et les travaux qui vont en perpétuer l'esprit se mèneront à l'écart du champ de l'IA.

[12] Les expressions citées sont extraites des transcriptions des archives de l'atelier : http://raysolomonoff.com/dartmouth/, consulté le 05/10/2018. Quant à la volonté de rompre avec la cybernétique, on ne peut être plus explicite que John McCarthy (1988) : « Quant à moi, une des raisons pour lesquelles j'ai inventé le terme "intelligence artificielle" était d'échapper à l'association avec la "cybernétique". Cette focalisation sur la rétroaction me semblait erronée, et je voulais éviter d'avoir à accepter Norbert Wiener comme gourou ou d'avoir à discuter avec lui. »

Un espace pour manipuler des symboles

La principale caractéristique de l'architecture des machines symboliques est de rompre le lien avec le monde et d'ouvrir un espace de raisonnement autonome au sein de leur calculateur. La configuration – dite « von Neumann » – des nouveaux ordinateurs qui se met en place dans les années 1950 instaure justement cet espace. Alors que l'ENIAC (1946) avait été conçu pour calculer des tables balistiques en « programmant » la machine dans le hardware, le projet de l'EDVAC (1952) sépare lui les opérations logiques effectuées sur des symboles (software) de la structure matérielle des machines (hardware) (von Neumann, 1945). Un espace propre est ainsi destiné au programme indépendamment du fonctionnement matériel de l'ordinateur. Celui-ci devient un « automate universel de calcul à programme centralisé » (Goldstine, 1972, pp. 198-199) et la programmation, indépendante des processus matériels, peut s'émanciper pour devenir un « travail de papier » selon la formule d'Alan Turing (2004, p. 21). Comme l'a montré Paul Edwards (1996), l'apparition de langages de programmation de haut niveau, proches du langage humain, ensuite compilés en langage machine sous forme de 0 et de 1, va permettre de désolidariser la machine physique de la machine symbolique ; l'intelligence artificielle peut désormais se penser comme une science de l'esprit dans la machine. L'une des premières contributions de l'IA à l'informatique tient justement à la conception de langage de programmation, dont le plus fameux est LISP, développé par John McCarthy en 1958, qui va pleinement être identifié à la recherche en IA en raison de ses capacités d'abstraction logique[13].

À peine ouvert au sein du calculateur, cet espace de programmation est disponible pour manipuler des symboles. L'IA naît la même année que les sciences cognitives (1956) et les deux domaines vont façonner ensemble le projet de doter les ordinateurs d'une capacité de raisonnement (Gardner, 1985). Contre la psychologie béhavioriste qui avait inspiré les « boîtes noires » adaptatives de la cybernétique, le projet des sciences cognitives est de donner à la machine des capacités abstraites et logiques. À la différence du connexionnisme, elles se désintéressent de la physiologie et du comportement humain pour ne porter attention qu'au raisonnement. La théorie computationnelle de l'esprit établit un dualisme en faisant l'hypothèse que les états mentaux peuvent être décrits à la fois sous forme matérielle comme un ensemble de traitements physiques des informations et sous forme symbolique comme des opérations, exécutables mécaniquement, de comparaison, de hiérarchisation ou d'inférence sur des significations (Andler, 2016). Cette hypothèse dite des « systèmes de symboles physiques » postule que l'esprit n'a pas un accès direct au monde, mais qu'il agit sur des représentations internes du monde pouvant être décrites et organisées sous la forme de symboles insérés dans des programmes.

Un monde « jouet »

Les fondateurs de l'IA ont tout fait pour se séparer des données du monde sensible et des comportements humains[14]. Le monde des machines symboliques est un décor de théâtre créé

[13] Une autre contribution de J. McCarthy au développement de l'IA est l'invention du temps partagé (*time sharing*), qui permet aux programmeurs d'interagir directement avec la machine et ses résultats, de lui parler, de la tester et de la rendre ainsi « intelligente » (Edwards, 1996).

[14] Comme le souligne J. Markoff (2015), toute l'histoire de l'informatique est tendue par une opposition entre ceux qui promeuvent l'intelligence dans la machine (*artificial intelligence – AI*) dont le SAIL, le laboratoire de J. McCarthy à Stanford est l'incarnation et l'obsession robotique l'emblème, et ceux qui cherchent à distribuer l'intelligence entre les humains et les

par la machine afin d'y projeter la syntaxe de leurs règles logiques : les jeux d'échecs ou de dames (Arthur Samuel), les théorèmes de géométrie (avec le Geometry Theorem Prover d'Herbert Gelertner), des décors de jeux vidéo. Les projets emblématiques de cette première vague de l'IA se caractérisent par l'invention d'espaces simplifiés de formes qu'il faut reconnaître et déplacer, comme les micromondes (MAC) de Marvin Minsky ou le célèbre langage SHLURDU de Terry Winograd. Tout comme l'espace restreint à quelques pièces et objets, dans lequel le robot Shakey est censé se déplacer, il s'agit d'espace fictif, « jouet »[15], dans lequel les objets peuvent facilement être associés à la syntaxe des règles qui sont calculées pour produire des comportements pertinents du système.

Si le *calculateur* projette son propre *monde*, c'est aussi parce qu'il a pour ambition de contenir lui-même son propre *horizon*. C'est en ce sens que cette IA a pu se revendiquer comme « forte » puisque les objectifs donnés au système lui sont propres et peuvent être déduits d'une sorte de raison incorporée aux inférences logiques effectuées par les modèles. Les très ingénieux langages inventés pour façonner la syntaxe de ces systèmes sont tous inférentiels. Ils organisent en étapes des opérations de traitement élémentaires transformant des entités dont chacune est une inférence d'un calcul correct (Andler, 1990, p. 100) : arbre de décision, chaîne de raisonnement intermédiaire, décomposition des buts et des sous-buts, analyse moyen/fin. L'horizon rationnel du calcul est enfermé dans la syntaxe du programme. La machine doit résoudre le problème, trouver la solution vraie ou correcte et prendre la décision satisfaisante[16]. Il n'est donc pas nécessaire de lui donner la bonne réponse (comme le feront les *exemples* des techniques d'apprentissage) puisque les règles doivent y conduire en suivant les inférences du calculateur. Syntaxe du raisonnement et sémantique des objets manipulés étant toutes les deux construites dans le calculateur, il est alors possible de confondre l'une et l'autre dans des raisonnements corrects et plus ou moins déterministes – mais au prix d'une conception artificialiste pour laquelle le monde « intelligent » est celui implémenté par le concepteur, un monde réglé, précis et explicite afin que la raison soit son horizon. Si, en chambre, ces machines réalisent quelques performances, elles vont rapidement se révéler aveugles et idiotes dès qu'un monde extérieur leur sera proposé.

Le premier hiver de l'IA

Au début des années 1970, l'IA entre dans son premier hiver et celui-ci va geler aussi bien les projets symboliques que connexionnistes. Les deux courants ont beaucoup trop promis et les résultats sont loin d'être au rendez-vous. Du côté connexionniste, le Perceptron de Frank Rosenblatt a souffert de la médiatisation à laquelle son promoteur – avec la complicité de l'US Navy – s'est livré sans retenue. Dans un concert de titres de presse enflammés par l'arrivée imminente de machines intelligentes, le *New York Times* annonce « l'embryon d'un ordinateur électronique dont la Marine espère qu'il marche, parle, voie, écrive, se reproduise

interfaces des machines (*Intelligence augmented – IA*) dont le laboratoire voisin de D. Engelbart sera le très fécond bastion et qui donnera naissance au courant *Human computer interaction* (HCI). Voir aussi Grudin (2009).

[15] Les micromondes, soulignent M. Minsky et S. Papert, sont « un pays enchanté dans lequel les choses sont si simplifiées que n'importe quelle assertion les concernant se révélerait totalement fausse si elle était transposée au monde réel » (Minsky et Papert, 1970). L'hypothèse proposée à travers cette réduction était qu'une représentation par un réseau de concepts abstraits au sein de micromondes pourrait ensuite être généralisée à un monde plus complet et plus riche. Les connexionnistes eux feront le raisonnement inverse : c'est une description au niveau le plus élémentaire des informations qui permet ensuite au réseau de généraliser.

[16] C'est par exemple la perspective mise en œuvre avec l'analyse fin/moyen du *General Public Solver* de Newell et Simon (1963).

lui-même et soit conscient de son existence »[17]. Mais c'est surtout du côté de l'IA symbolique que, Herbert Simon et Marvin Minsky en tête, les prophéties et les annonces exagérées sont rapidement démenties. Intoxiqués par les promesses des chercheurs, les militaires et la DARPA pensant disposer rapidement de traducteurs de textes russes, de robots infiltrés dans les lignes ennemies ou de systèmes de commande vocale pour les pilotes de tank et d'avion, découvrent que les systèmes « intelligents » annoncés ne sont que des jeux en chambre. En 1966, signe avant-coureur, le Conseil national de la recherche coupe les crédits en traduction automatique, décision qui va lancer une cascade de désinvestissements de la part des soutiens financiers et académiques de l'IA. Au début des années 1970, le projet *micromondes* de Minsky et Papert au MIT est à la peine et perd ses soutiens. À Stanford, c'est le robot Shakey qui ne reçoit plus de financement militaire et le programme de reconnaissance de la parole SUR de la DARPA dont bénéficiait Carnegie Mellon est arrêté brutalement. En Angleterre, le très critique *Lighthill report*, en 1973, va lui aussi convaincre d'arrêter les financements publics de l'IA (Crevier, 1997, pp. 133-143).

Avec la crise des financements, c'est le projet même d'une modélisation logique du raisonnement qui est mis à mal par des critiques qui se rendent alors de plus en plus visibles. La Rand commande en 1965 au philosophe Hubert Dreyfus un rapport sur l'IA qu'il titre « L'alchimie et l'IA » et qui va mettre en place une vigoureuse argumentation qui sera développée dans la première édition d'un ouvrage à succès, *What Computer can't do* (Dreyfus, 1972). Âpre et violente, la polémique entre l'establishment de l'IA et Hubert Dreyfus va fragiliser considérablement l'idée que des règles de raisonnement puissent rendre des machines « intelligentes ». L'explicitation de règles logiques manque complètement les formes corporelles, situées, implicites, incarnées, collectives et contextuelles de la perception, de l'orientation et des décisions des comportements humains[18]. La critique sera aussi portée par la première génération de « renégats » qui vont devenir de redoutables contempteurs d'espérances qu'ils ont eux-mêmes portées comme Joseph Weizenbaum (1976), l'initiateur d'Eliza ou Terry Winograd, le concepteur déçu de SHRDLU (Winograd et Flores, 1986). Les machines « intelligentes » raisonnent avec de belles règles logiques, une syntaxe déterministe, et des objectifs rationnels, mais leur monde n'existe pas.

LA DEUXIÈME VAGUE DE L'IA : UN MONDE D'EXPERTS

L'IA connaît cependant un deuxième printemps dans les années 1980 en proposant sous le nom de « système expert »[19] une importante révision de l'architecture des machines symboliques. Cette renaissance a été rendue possible par l'accès à des calculateurs plus puissants permettant de faire entrer des informations beaucoup plus nombreuses dans la mémoire des ordinateurs. Aux mondes « jouets » peut alors être substitué un répertoire de « connaissances spécialisées » prélevées dans le savoir d'experts[20]. Les artefacts de la deuxième

[17] « Electronic "Brain" Teaches Itself », *New York Times*, 13 juillet 1958.

[18] À la suite de l'ouvrage de H. Dreyfus, et souvent au contact des sciences humaines et sociales, un courant très productif de critiques de l'IA va se développer, autour de la critique wittgensteinienne de la règle. Il donnera lieu aux travaux sur la distribution de l'intelligence dans l'espace (Collins), la forme collective de la cognition (Brooks) ou l'inscription corporelle de l'esprit (Varela).

[19] Les autres qualifications des machines intelligentes de la seconde vague de l'IA sont : « intelligent knowledge-based systems », « knowledge engineering », « office automation » ou « multiagent systems ».

[20] En 1967, dans un discours à Carnegie prononcé devant A. Newell et H. Simon, E. Feigenbaum lance un défi à ses anciens professeurs : « Vous travaillez sur des problèmes jouets (*toy problems*). Les échecs et la logique sont des problèmes jouets. Si

génération de l'IA interagissent avec un monde extérieur qui n'a pas été conçu et façonné par les programmateurs : il est désormais composé de connaissances qu'il faut aller chercher auprès de spécialistes de différents domaines, transformer en un ensemble de propositions déclaratives, formuler dans un langage le plus naturel possible (Winograd, 1972) afin que des utilisateurs puissent interagir avec elles en leur posant des questions (Goldstein et Papert, 1977). Cette extériorité du *monde* à calculer conduit à modifier la structure des machines symboliques en séparant le « moteur d'inférence » qui en constitue désormais le *calculateur* et une série de *mondes* possibles appelés « systèmes de production » selon la terminologie proposée par Edward Feigenbaum pour DENDRAL, le premier système expert permettant d'identifier les composants chimiques des matériaux. Les données qui nourrissent ces bases de connaissances consistent en de longues listes facilement modifiables et révisables de règles du type « SI... ALORS » (par exemple : « SI FIÈVRE, ALORS [CHERCHER INFECTION] ») qui sont dissociées du mécanisme permettant de décider quand et comment appliquer la règle (moteur d'inférence). MYCIN, la première réalisation d'une base de connaissances de 600 règles destinées à diagnostiquer les maladies infectieuses du sang sera le point de départ, dans les années 1980, du développement d'une ingénierie des connaissances qui s'applique essentiellement à des contextes scientifiques et industriels : XCON (1980) aide les clients des ordinateurs de DEC à les configurer, DELTA (1984) identifie les pannes des locomotives, PROSPECTOR repère des gisements géologiques, etc. (Crevier, 1997, p. 233 et suiv.). Les grandes industries développent des équipes d'IA au sein de leur organisation, les chercheurs se lancent dans l'aventure industrielle, les investisseurs se précipitent sur ce nouveau marché, des entreprises croissent à une vitesse exceptionnelle (Teknowledge, Intellicorp, Inference) – toujours avec un soutien fidèle de l'ARPA (Roland et Shiman, 2002) – et les médias s'emparent du phénomène pour, à nouveau, annoncer l'arrivée imminente des « machines intelligentes » (Waldrop, 1987).

Des cathédrales de règles

Face aux critiques reprochant au computationnalisme rigide du premier âge d'inventer un univers abstrait sans lien réaliste avec le monde, les recherches en IA vont engager, par le haut, un processus de complétion, d'intellection et d'abstraction des systèmes conceptuels destinés à manipuler les entités de ces nouvelles bases de connaissance. Le projet symbolique renforce alors son ambition rationalisatrice par une surenchère modélisatrice afin d'intégrer la variété des contextes, les imperfections du raisonnement et la pluralité des heuristiques et se rapprocher ainsi, à travers la médiation des experts, du monde des utilisateurs. Cet investissement dans la programmation du calculateur se caractérise par une relaxation des opérateurs logiques (syntaxe) et une densification des réseaux conceptuels permettant de représenter les connaissances (sémantique). Le mouvement qui s'observe au sein des recherches en IA cherche à désunifier le mécanisme central, générique et déterministe du raisonnement computationnel pour pluraliser, décentraliser et probabiliser les opérations effectuées sur les connaissances. Empruntant notamment aux discussions sur la modularité de l'esprit (Fodor, 1983), les systèmes implémentés dans les calculateurs décomposent le processus de raisonnement en briques élémentaires, des « agents » en interactions, qui de façon autonome peuvent avoir des manières différentes de mobiliser des

vous parvenez à les résoudre, vous ne faites que résoudre des problèmes jouets. Et c'est tout ce que vous aurez fait. Sortez dans le monde réel et essayez de résoudre des problèmes du monde réel ! » (Feigenbaum et McCorduck, 1983, p. 63).

connaissances et d'en inférer des conséquences[21]. Aussi est-ce dans l'organisation sémantique des significations des heuristiques issues des bases de connaissances qu'ont été conçues les principales innovations de cette deuxième vague de l'IA symbolique. Elles mobilisent des langages (PROLOG, MICROPLANNER, CYCL) et des constructions intellectuelles d'une rare sophistication : principe des listes, notion de « dépendance conceptuelle » élaborée par Robert Schank, réseaux sémantiques de Ross Quillian... Le chef-d'œuvre inabouti de ces multiples initiatives sera Cyc, une entreprise d'ontologie générale des connaissances de sens commun menée par Donald Lenat bâtie sur une architecture de « prédicats fondamentaux », de « fonctions de vérité » et de « micro-théories » qui, au sein de la communauté de l'IA, suscitera l'admiration de tous et ne sera utilisée par personne.

Le volume croissant des connaissances en entrée et la complexification des réseaux de concepts destinés à les manipuler sont à l'origine d'un autre déplacement important : les règles de raisonnement deviennent conditionnelles et peuvent être probabilisées. Face à l'approche rationnelle et logique incarnée par John McCarthy, Marvin Minsky et Samuel Papert défendent dès les années 1970 l'idée que « la dichotomie correct/mauvais est trop rigide. En traitant des heuristiques plutôt que de la logique, la catégorie vrai/faux est moins importante que fructueux/stérile. Naturellement, l'objectif final doit être de trouver une véritable conclusion. Mais, que les logiciens ou les puristes le veuillent ou non, le chemin de la vérité passe principalement par des approximations, des simplifications et des intuitions plausibles qui sont en fait fausses lorsqu'elles sont prises littéralement » (Minsky et Papert, 1970, p. 41). Parmi les milliers de règles formulées par les experts, il est possible à partir d'une prémisse fixée (SI...), d'établir une probabilité relative au fait que la deuxième proposition (ALORS...) a un pourcentage de chance d'être vraie. La probabilisation des règles de connaissance permet de relaxer la forme déterministe du raisonnement inférentiel qui a connu son heure de gloire dans les premiers temps de l'IA. En devenant plus réalistes, plus diverses et contradictoires, les connaissances qui entrent dans les machines à prédire y font aussi pénétrer les probabilités (Nilsson, 2010, p. 475). Si le couple « fructueux/stérile » se substitue à « vrai/faux », l'*horizon* qui donne son objectif au calculateur apparaît alors moins comme une vérité logique que comme une estimation du caractère correct, pertinent ou vraisemblable des réponses données par le système. Mais cette estimation ne peut alors plus être prise en charge, de façon immanente, par les règles du calculateur. Elle doit être externalisée vers le monde constitué par les experts qui sont mobilisés pour fournir des exemples et des contre-exemples aux mécanismes d'apprentissage artificiel (*machine learning*) [22].

Avec la probabilisation des inférences, ces techniques pénètrent de plus en plus le champ de l'IA afin de réaliser des tâches qu'il est devenu impossible, pour un programmeur, de réaliser « à la main » (Carbonnell *et al.*, 1983). À la suite des travaux de Tom Mitchell (1977), les méthodes d'apprentissage peuvent être décrites comme une solution statistique pour trouver le meilleur modèle au sein d'un *espace d'hypothèses* – ou de « versions » – générées

[21] La théorie des « cadres » développée par M. Minsky (1975) sera très influente dans ce processus et donnera lieu à une théorisation globale dans *La société de l'esprit* (1986).

[22] Pour les tenants de la logique, à l'instar d'A. Newell, une telle position est une hérésie : « Vous avez des experts qui travaillent pour vous et quand vous avez un problème, vous décidez quel expert doit être appelé pour résoudre le problème. Ce n'est pas de l'IA » (McCorduck, 1979, p. 229).

automatiquement par le calculateur. Avec les systèmes experts, cet espace d'hypothèses est fortement structuré par la nature des données en entrée – i.e. les « connaissances ». Le mécanisme d'apprentissage « explore » les multiples versions de modèles produites par le calculateur afin de rechercher une hypothèse cohérente en déployant des inférences logiques pour construire des raisonnements (généralisation de concept, subsomption, déduction inverse). Les méthodes statistiques pour éliminer les hypothèses candidates vont ainsi mûrir et se développer en produisant des raisonnements de type inférentiel, comme les arbres de décision (qui, par la suite, donneront naissance aux forêts aléatoires (*random forests*)), les techniques dites « *divide and conquer* » ou les réseaux bayésiens qui permettent d'ordonner des dépendances entre variables avec un formalisme causaliste (Domingos, 2015). Même automatisée, la découverte automatique d'une fonction cible conserve l'idée que les modèles sont des hypothèses et que si la machine n'applique certes plus un raisonnement déductif, elle choisit le meilleur raisonnement possible parmi un ensemble de raisonnements potentiels. Cependant, à partir du début des années 1990, le changement de la nature des données qui compose le monde en entrée du calculateur, va conduire à un déplacement au sein même du champ de l'apprentissage artificiel. Les données sont plus nombreuses, elles ne sont plus organisées sous forme de variables étiquetées, de concepts interdépendants et bientôt elles vont perdre toute intelligibilité pour devenir des vecteurs numériques (*infra*). Ne portant plus de structure, les données ne peuvent plus être rassemblées que sous la forme d'un voisinage statistique. On assiste alors à une bascule au sein du champ de l'apprentissage artificiel passant de méthodes « par exploration » à des méthodes « par optimisation » (Cornuéjols *et al.*, 2018, p. 22) qui vont faire s'effondrer les cathédrales de règles au profit de calculs statistiques massifs.

En élargissant de plus en plus le volume et le réalisme des données à calculer, le mécanisme inductif change de direction à l'intérieur du calculateur. Si les données ne donnent plus d'informations sur les relations qu'elles ont entre elles (catégories, dépendances entre variables, réseaux conceptuels) alors, pour identifier la fonction cible, le mécanisme inductif va s'appuyer sur le critère d'optimisation final pour faire la bonne partition (Cornuéjols *et al.*, 2018, p. 22). La transformation de la composition du monde à apprendre conduit les chercheurs à modifier la méthode inductive mise en œuvre et, ce faisant, à proposer une architecture toute différente des machines à prédire. On verra ce déplacement s'accélérer avec les réseaux de neurones (*infra*), mais le tournant avait déjà été préparé au sein du monde de l'apprentissage artificiel. Les données étant de moins en moins « symboliques », le mécanisme inductif ne cherche plus le modèle dans la structure des données initiales, mais dans le facteur d'optimisation (Mazières, 2016). L'horizon du calcul n'est plus interne au calculateur, mais est une valeur que le monde lui donne de l'extérieur – et qui est souvent très « humaine », comme va en témoigner tout ce travail manuel d'étiquetage des données : cette image contient-elle (ou pas) un rhinocéros ? Cet utilisateur a-t-il cliqué (ou pas) sur tel lien ? La réponse (le critère d'optimisation) doit être entrée dans le calculateur avec les données pour que celui-ci découvre un « modèle » adéquat. Les nouvelles méthodes d'apprentissage (SVM, réseau de neurones) vont ainsi se révéler plus efficaces tout en devenant inintelligibles, comme le soulignera l'inventeur des arbres de décision, Léo Breiman (2001), dans un article provocateur sur les deux cultures de la modélisation statistique.

Les sublimes cathédrales échafaudées par les bâtisseurs de systèmes experts n'ont pas tenu leurs promesses. Elles sont rapidement apparues d'une complexité extrême et très limitées dans leurs performances. Le marché très dynamique qui s'était constitué au milieu des années 1980 s'est brutalement effondré et les prometteuses entreprises de l'IA ont fait faillite, notamment parce que, pour vendre les systèmes experts, elles vendaient aussi des stations de travail spécialisées, dites « machine LISP », à un prix exorbitant au moment où le marché des PC prenait son essor (Markoff, 2015, p. 138 et suiv.). La diminution du coût et l'augmentation des capacités de calcul dans les années 1980 ouvrent l'accès à des calculateurs puissants aux courants hétérodoxes et déviants qui avaient été mis à l'écart par le monopole du courant symbolique sur le financement des grands projets informatiques (Fleck, 1987, p. 153). Le contrôle du petit cercle d'universités influentes sur la définition « symbolique » de l'IA s'affaiblit d'autant plus, que dans les domaines de la synthèse vocale, de la reconnaissance de formes et d'autres secteurs où les données numériques sont importantes, les systèmes experts ne produisent guère de résultats. Au début des années 1990, l'IA symbolique est si affaiblie que l'usage de ce terme disparaît quasiment du vocabulaire de la recherche. Compléter *ad infinitum* des répertoires de règles explicites afin qu'elles épousent les mille subtilités de la perception, du langage et du raisonnement humain est de plus en plus apparu comme une tâche impossible, déraisonnable et inefficace (Collins, 1992 ; Dreyfus, 2007).

LES REPRÉSENTATIONS DISTRIBUÉES DU *DEEP LEARNING*

C'est dans ce contexte que, sortant de la phase de relégation débutée à la fin des années 1960, les approches connexionnistes débutent dans les années 1980 et 1990 une période de renaissance d'une très grande créativité théorique et algorithmique. À la suite d'une rencontre organisée en juin 1979 à La Jolla (Californie) par Geoff Hinton et James Anderson, un groupe de recherche interdisciplinaire rassemblant biologistes, physiciens et informaticiens propose de se pencher à nouveau sur le caractère massivement distribué et parallèle des processus mentaux pour en faire une alternative au cognitivisme classique. Ce groupe acquiert une réelle visibilité en 1986 avec la parution de deux volumes de travaux réunis sous le nom de *Parallel Distributed Processing* (PDP), terme retenu pour échapper à la mauvaise réputation de celui de « connexionnisme » (Rumelhart *et al.*, 1986b). Contre les approches séquentielles de l'ordinateur et du raisonnement symbolique, les travaux du PDP explorent les microstructures de la cognition en exploitant à nouveau la métaphore des neurones pour dessiner un contre-modèle aux propriétés originales : des unités élémentaires sont liées entre elles par un vaste réseau de connexions ; les connaissances ne sont pas statiquement stockées, mais se logent dans la force des connexions entre unités ; celles-ci communiquent les unes avec les autres par un mécanisme d'activation binaire (« la monnaie de notre système n'est pas le symbole, mais l'excitation et l'inhibition », p. 132) ; ces activations se font tout le temps, de façon parallèle et non en suivant les étapes d'un processus ; il n'y a pas de centre de commandement central sur les flux ; une sous-routine n'engendre pas le comportement d'une autre, mais les sous-systèmes modulent le comportement d'autres sous-systèmes en produisant des contraintes qui sont factorisées dans les calculs ; les opérations que réalise la machine s'apparentent à un système de relaxation dans lequel le calcul procède itérativement à des approximations pour satisfaire un grand nombre de contraintes faibles (« le système doit plus être pensé comme *installant*

une solution que comme *calculant* une solution », p. 135). Le dispositif que conçoivent les connexionnistes forme bien des représentations internes, et celles-ci peuvent être de haut niveau, mais elles sont « sous-symboliques » (*sub symbolic*), statistiques et distribuées (Smolensky, 1988). Comme en témoigne ce rapide résumé, l'approche connexionniste n'est pas une simple méthode, mais une très ambitieuse construction intellectuelle destinée à renverser le cognitivisme computationnel.

> « Au premier temps, dans les années 1950, des gens comme von Neumann et Turing ne croyaient pas à l'IA symbolique, *explique Geoff Hinton*. Ils étaient beaucoup plus inspirés par le cerveau. Malheureusement, ils sont tous les deux morts beaucoup trop jeunes, et leur voix n'a pas été entendue. Au début de l'IA, les gens étaient absolument convaincus que la représentation de l'intelligence dont nous avons besoin était une forme d'expression symbolique, pas tout à fait de la logique, mais quelque chose comme de la logique : l'essence de l'intelligence était le raisonnement. Ce qui arrive aujourd'hui, c'est un point de vue complètement différent, à savoir qu'une pensée n'est qu'un grand vecteur d'activité neuronale. Je crois que les gens qui pensaient que les pensées étaient des expressions symboliques ont fait une énorme erreur. Ce qui entre, c'est une chaîne de mots [*string of words*], et ce qui sort, c'est une chaîne de mots. Et à cause de cela, les chaînes de mots sont apparues comme la manière évidente de représenter les choses. Ils ont donc pensé que ce qui devait se trouver entre les deux était une chaîne de mots, ou quelque chose comme une chaîne de mots. Je pense que ce qui se trouve entre les deux n'a rien à voir avec une chaîne de mots ! [...] Les pensées ne sont que ces grands vecteurs et les grands vecteurs ont un pouvoir causal. Ils provoquent d'autres grands vecteurs, et c'est tout à fait différent de la vision standard de l'IA[23]. »

Si ces références épistémiques se sont aujourd'hui atténuées chez les nouveaux utilisateurs pragmatiques des réseaux de neurones, qui n'ont pas connu les exclusions et les moqueries dont leurs aînés ont été victimes, elles vont constituer un ressort constant de la poursuite opiniâtre du projet connexionniste. Ce qu'il faut insérer entre les chaînes de mots qui entrent et celles qui sortent n'est pas un modèle programmé par un esprit logicien, mais un réseau d'entités élémentaires qui adapte ses coefficients aux entrées et aux sorties. Autant que possible, il est nécessaire qu'il le « fasse tout seul », et cela justement requiert de nombreux artifices.

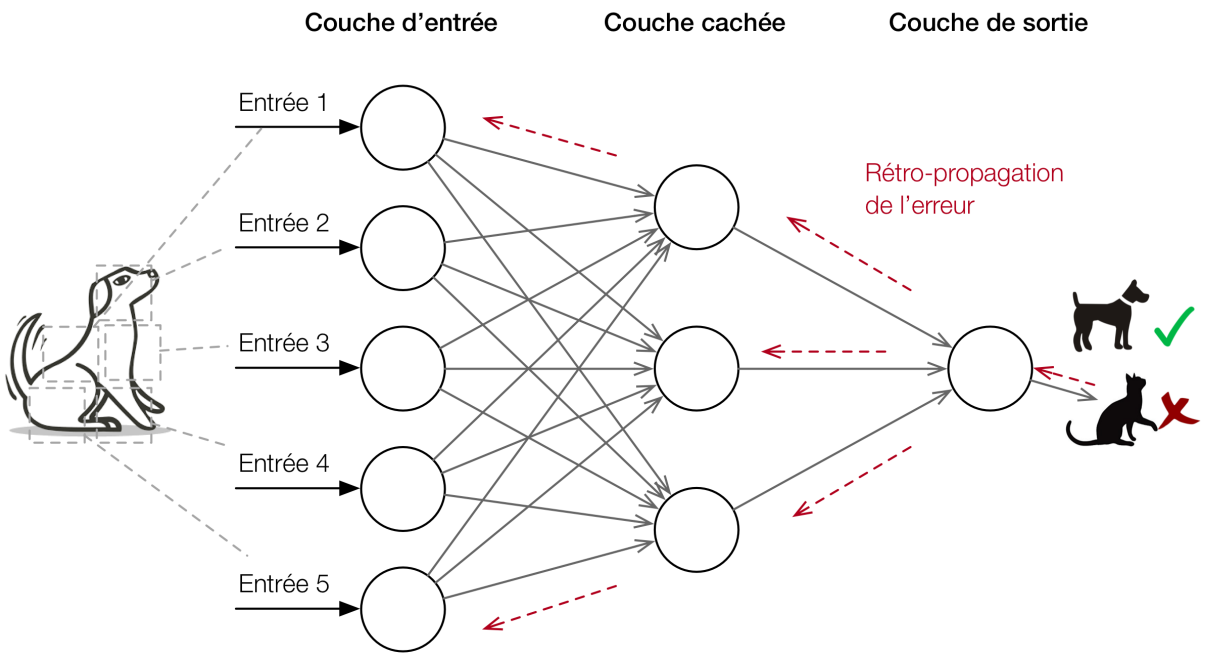
La reconfiguration algorithmique du connexionnisme

Inspirée par le travail de John Hopfield qui proposa une révision du modèle du Perceptron en donnant à chaque neurone la possibilité de mettre à jour ses valeurs de façon indépendante, le physicien Terry Sejnowski et le psychologue anglais Geoff Hinton vont développer au début des années 1980 de nouvelles architectures multi-couches pour les réseaux de neurones (appelé machine de Boltzmann) et concevoir Nettalk, un système à trois couches de neurones et 18 000 synapses qui parvint à transformer des textes en phrases

[23] Hinton G., « Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton », YouTube, 8 août 2017 (37'20 et suivante).

vocalisées. Mais le véritable point de bascule de ce renouveau est la conception d'un algorithme, la rétropropagation de gradient stochastique (la « backprop »), qui permet de calculer le poids des coefficients (Rumelhart *et al.*, 1986a). Dépassant la critique de Minsky et Papert (1969), les auteurs montrent que, lorsqu'on le dote de plusieurs couches, il est possible d'entraîner de façon simple un réseau de neurones ; les couches additionnelles de neurones permettent d'apprendre des fonctions non linéaires. L'algorithme fonctionne en prenant la dérivée de la fonction de perte du réseau et en « propage » l'erreur pour corriger les coefficients dans les couches basses du réseau[24] – dans un esprit proche des machines cybernétiques, l'erreur en sortie est « propagée » vers les entrées (figure 5).

Figure 5. Fonctionnement d'un réseau de neurones simple



Fortes d'un algorithme à vocation générale permettant d'optimiser tout type de réseaux de neurones, les années 1980 et 1990 constituent une remarquable période d'inventivité et marquent le renouveau du connexionnisme. Un des premiers succès est l'application faite par Yann Lecun sur la reconnaissance des codes postaux effectuée à AT&T Bell Labs (Lecun *et al.*, 1989) qui « invente » la technique de la convolution. En utilisant une base de données de l'*US Postal Service*, il parvient à entraîner un réseau multicouche pour reconnaître les chiffres du code postal inscrit sur les colis. Le succès de son approche devient une des premières utilisations industrielles généralisées des réseaux de neurones dans les secteurs bancaire (vérification des montants des chèques) et postal. Suivront ensuite toute une série de propositions pour accueillir un plus grand nombre de couches cachées, complexifier la carte des connexions (encodeurs), diversifier les fonctions d'optimisation (RELU), intégrer de la mémoire dans les couches du réseau (*réseaux récurrents et LSTM*), mixer selon les parties du réseau apprentissage non supervisé et supervisé (*beliefs network),* etc. (Kurenkov, 2015). De façon très créative, de nombreuses architectures câblant différemment les relations entre les neurones sont alors testées pour en explorer les propriétés.

[24] Il existe un débat d'antériorité sur l'algorithme de la « backprop ». Cette méthode a été formulée et utilisée à maintes reprises avant la publication de cet article, notamment par Linnainmaa en 1970, Werbos en 1974 et Lecun en 1985.

« C'est pas convexe, mais c'est plus efficace ! »

Si ces algorithmes jettent les bases de la plupart des approches qui caractérisent aujourd'hui le *deep learning*, leur invention n'est pas immédiatement couronnée de succès. De 1995 à 2007, les soutiens institutionnels sont devenus très rares, les papiers sont refusés dans les conférences et les résultats obtenus restent limités. « *Ils ont vécu un hiver colossal, raconte un chercheur en computer vision, la réalité, c'est que, à l'époque, personne n'arrivait à faire marcher ces machines. Tu avais cinq labos dans le monde qui savaient, nous on n'arrivait pas à les entraîner*[25]. » Les chercheurs qui maintiennent ces techniques autour de Geoff Hinton, Yann LeCun et Yoshua Bengio constituent un petit groupe isolé, mais solidaire, qui aura pour seul et principal soutien le Canadian Institute for Advanced Research (CIFAR). Leur situation est rendue d'autant plus difficile qu'une technique d'apprentissage originale voit le jour en 1992, les machines à vecteur de support (SVM) – aussi appelées « méthodes à noyaux » –, qui se révèlent très performantes sur de petits *datasets*. Déjà au ban de l'intelligence artificielle, les connexionnistes se retrouvent alors aux marges de la communauté de l'apprentissage artificiel.

> « À cette époque, si tu disais que tu faisais un réseau de neurones, tu ne pouvais pas passer un papier. Jusqu'en 2010, c'était comme ça, un truc de has been. Je me souviens, LeCun, on l'avait dans le labo en prof invité et il fallait se dévouer pour aller manger avec lui. Personne ne voulait y aller. C'était la poisse, je te jure. Il pleurait, ses papiers étaient refusés à CVPR, ses trucs n'étaient pas à la mode, c'était pas sexy. Donc les gars allaient vers les trucs à la mode. Ils allaient vers les noyaux, SVM machin. Alors Lecun il disait : "J'ai un réseau de neurones avec 10 couches, il fait pareil." Et on lui disait : "Ah bon t'es sûr ? Et qu'est-ce qu'il y a de nouveau ?" Parce que, une fois que t'as posé un réseau de neurones, bon d'accord cette fois il a 10 couches, mais il marche pas mieux que l'autre. C'était pourri ! Alors il disait : "Mais oui, mais il y a pas assez de données !"[26]. »

Parmi les reproches opposés aux rares promoteurs des réseaux de neurones, un argument revient alors sans cesse.

> « Ils [*les tenants des SVM*] disaient toujours : "C'est pas convexe votre truc, c'est juste un tour de main !", raconte un autre chercheur. Ils n'avaient que cela à la bouche. Nous, on présentait des papiers et, eux : "C'est pas convexe !". C'étaient des matheux, obsédés par l'optimisation, qui n'avaient jamais vu que ça dans leur vie ! Pendant des années, on a eu ça. Alors que, nous, on s'en foutait complètement[27]. »

En raison de leur non-linéarité constitutive[28], les réseaux de neurones ne peuvent pas garantir que lors de la phase d'optimisation de la fonction de perte, le minimum global ait

[25] Interview V., chercheur en computer vision, 12 mars 2018.
[26] *Ibid.*
[27] Interview F., chercheur en informatique, un des pionniers du *deep learning* en France, 20 juillet 2018.
[28] L'originalité des réseaux de neurones est que la fonction d'activation des neurones crée des discontinuités qui produisent des transformations non linéaires : une sortie ne peut pas être reproduite par une combinaison linéaire des entrées.

été trouvé ; il se peut très bien qu'elle converge vers un minimum local ou un plateau[29]. Dans les années 2005-2008, une véritable politique de reconquête est initiée par le petit groupe de la « conspiration des neurones » (Markoff, 2015, p. 150) pour convaincre la communauté du *machine learning* qu'elle est victime d'une épidémie de « convexitis » (LeCun, 2007). Alors qu'en 2007 leurs papiers avaient été refusés à NIPS, ils organisent une session satellite, un *off*, en transportant par cars les participants à l'hôtel Hyatt de Vancouver pour y défendre une approche que la domination des SVM donne alors comme archaïque et alchimique. Yann LeCun porte le fer en titrant son exposé : « Qui a peur des fonctions non convexes ? » Après avoir présenté plusieurs résultats montrant que les réseaux de neurones étaient plus performants que les SVM, il soutient qu'un attachement trop étroit à des réquisits théoriques issus de modèles linéarisés empêche d'imaginer des architectures de calcul innovantes et de porter attention à d'autres méthodes d'optimisation. Certes, la technique très simple de la descente de gradient stochastique ne garantit pas la convergence vers un minimum global, mais « quand des preuves empiriques suggèrent un fait pour lequel vous n'avez pas de garanties théoriques, cela veut juste dire que la théorie est inadaptée [...], si pour cela, vous avez dû jeter la convexité par la fenêtre, c'est très bien ! » (LeCun, 2017, 11'19).

> « Les créatifs, c'est toujours des fous furieux, *commente un participant à cette controverse*. Au départ, cette bande-là, les créatifs, était d'une grande effervescence. Et puis après arrivent les mecs qui ne sont pas dans l'IA, qui viennent des mathématiques et qui crachent sur la descente de gradient pour te balancer leurs trucs : mon théorème est plus beau que le tien. En optimisation, les gens ont passé, je ne sais pas, dix ans à rechercher une méthode convexe plus rusée et à faire des choses sophistiquées, mais très coûteuses [*en capacité de calcul*]. C'est pas inintéressant, mais c'est l'assèchement total, des milliers de papiers là-dessus et quand la grande vague des données est arrivée, plouf, tous leurs machins ne marchaient pas[30]! »

Transformer le monde en vecteurs

Les connexionnistes vont ainsi déplacer la controverse scientifique sur la convexité en demandant aux nouveaux flux de données qui se présentent aux portes des laboratoires d'arbitrer le choix de la meilleure méthode de calcul. L'architecture des machines à prédire va être transformée pour affronter les *big data*. Celles-ci ne ressemblent en rien aux petits *datasets* calibrés et très artificiels des traditionnelles compétitions entre chercheurs. Car, pendant ce débat, l'informatisation de la société et les développements des services du web ont fait émerger de nouveaux problèmes d'ingénierie à base de grands volumes de données comme la détection de spams, les techniques de filtrage collaboratif utilisées pour la recommandation, la prédiction de stock, la recherche d'information ou l'analyse des réseaux sociaux. Dans ce contexte industriel, les méthodes statistiques de la nouvelle science des données (*datascience*) empruntent et développent des techniques d'apprentissage artificiel (méthode bayésienne, arbre de décision, forêt aléatoire, etc.) sans se préoccuper de se

[29] La propriété qui assure la réputation des SVM est de proposer un système non linéaire qui peut être régularisé pour garantir la convexité (Boser et al., 1992).
[30] Interview F., un des pionniers du deep learning en France, 20 juillet 2018.

positionner par rapport aux enjeux de l'IA (Dagiral et Parasie, 2017). En revanche, il apparaît clairement que face au volume et à l'hétérogénéité des caractéristiques des données, plutôt que des techniques « confirmatoires », il était nécessaire d'utiliser des méthodes plus « exploratoires » et inductives (Tuckey, 1962). Aussi est-ce au contact des acteurs industriels (AT&T originellement, puis Google, Facebook et Baidu) que les conspirateurs des réseaux de neurones vont aller à la rencontre de problématiques, de capacités de calcul et de jeux de données leur permettant de démontrer les potentialités de leurs machines et d'imposer leur vue dans la controverse scientifique. Ils vont y introduire un nouvel arbitre : l'efficacité des prédictions lorsqu'elles sont appliquées, cette fois, au monde « réel ».

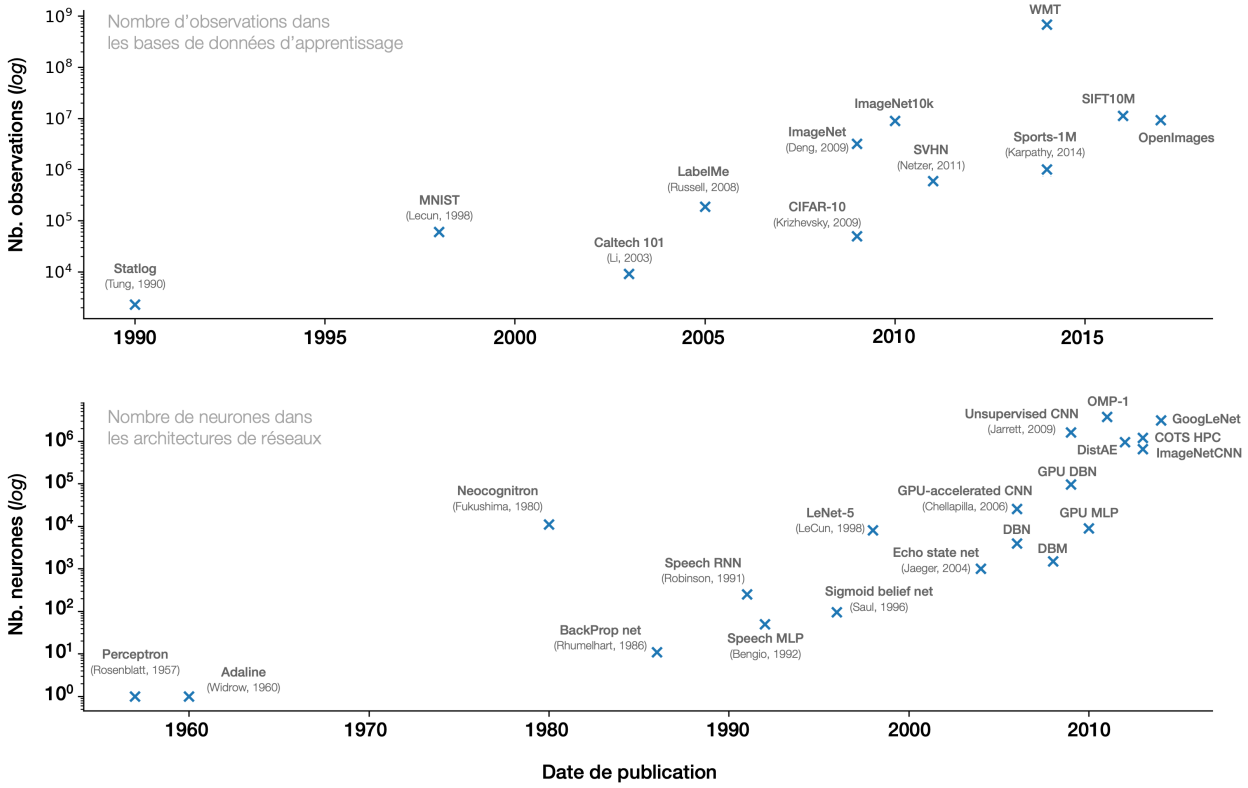
Les néo-connexionnistes imposent d'abord leurs propres termes dans le débat. Il faut, expliquent-ils, distinguer la « largeur » des architectures « minces » (*shallow*) des SVM de la « profondeur » (le terme de *deep learning* est ainsi forgé en 2006 par Geoff Hinton) des architectures en couches de neurones. Ainsi peuvent-ils démontrer que la profondeur est préférable à la largeur : seule la première est calculable lorsque données et dimensions augmentent et parvient à capturer la diversité des caractéristiques des données. Tout convexes soient-ils, les SVM ne donnent pas de bons résultats sur les jeux de données volumineux : les dimensions augmentent trop vite et deviennent incalculables, les mauvais exemples provoquent des perturbations considérables sur la prédiction, la solution consistant à linéariser une méthode non linéaire fait perdre au système sa capacité à apprendre des représentations complexes (Bengio et LeCun, 2007). Les croisés du connexionnisme parviennent ainsi à convaincre qu'il est préférable de sacrifier l'intelligibilité du calculateur, et une optimisation rigoureusement contrôlée, à une meilleure perception de la complexité des dimensions présentes dans ces nouvelles données. Quand le volume des données d'entraînement augmente considérablement, il existe beaucoup de minimums locaux, mais il se forme assez de redondances et de symétries pour que les représentations apprises par le réseau soient robustes et tolérantes aux erreurs dans les données d'apprentissage. Au cœur de ce débat tendu avec la communauté du *machine learning*, un sous-entendu est omniprésent : il n'y a que dans les laboratoires que les modèles sont linéaires, le monde, le « vrai monde », celui des données produites par la numérisation des images, des sons, des paroles et des textes, lui, est non linéaire. Il est bruité, l'information y est redondante, les flux de données ne sont pas catégorisés derrière des attributs de variables homogènes, claires et construites de façon intelligible, les exemples sont parfois faux. « Une IA, écrivent Yoshua Bengio *et al.* (2013), doit comprendre fondamentalement le monde qui nous entoure, et nous soutenons que cela ne peut être réalisé que si elle peut apprendre à identifier et à démêler les facteurs explicatifs sous-jacents cachés dans le milieu observé des données sensorielles de bas niveau. » C'est pourquoi une architecture « profonde » est plus calculable et plus « expressive » qu'une architecture « mince » (Lecun et Bengio, 2007). Diminuer l'intelligibilité du calculateur pour capturer plus de complexité du monde, cette polémique sur la convexité montre bien que, loin d'un empiricisme naïf, la production des machines inductives est le résultat d'un intense travail pour convaincre de la nécessité de reformuler de façon essentielle la relation entre le calculateur et le monde.

Aussi, pour que les données fassent basculer le débat scientifique, a-t-il été nécessaire d'augmenter de façon absolument radicale le volume des *datasets* de recherche. Dans l'article de 1988 sur la reconnaissance de caractère, Yann LeCun a utilisé une base de 9 298 chiffres

manuscrits de codes postaux. La base qui depuis 2012 sert à la reconnaissance des caractères (MNIST) contient, elle, 60 000 données étiquetées d'images en noir et blanc de 28 pixels de côté. Elle a permis de démontrer l'efficacité des réseaux de neurones, mais pas d'emporter l'adhésion face à d'autres techniques comme les SVM. Aussi les communautés scientifiques vont-elles profiter du web pour produire des *datasets* beaucoup plus volumineux et les construire explicitement pour des tâches d'apprentissage en créant des couples entrées/sorties. Cette capture systématique, extensive et la plus élémentaire possible de données numériques, permet de donner plus de sens à la formule d'Hubert Dreyfus soutenant que « le meilleur modèle du monde est le monde lui-même » (Dreyfus, 2007, p. 1140). Comme le défendaient depuis longtemps les approches hétérodoxes critiques de l'IA représentationnelle, les représentations sont dans les données du monde et non pas internes au calculateur (Brooks, 1988). La fabrication d'ImageNet, le *dataset* utilisé lors du challenge présenté au début de cet article, qui a été initiée par Lee Fei Fei (Deng *et al.*, 2009), est à cet égard exemplaire. Cette base de données comprend aujourd'hui 14 millions d'images dont les éléments ont été annotés manuellement en 21 841 catégories en s'appuyant sur la structure hiérarchique d'une autre base de données classique en Traitement Automatique de la Langue, Wordnet (Miller, 1995). Pour mener à bien cet immense travail de qualification des éléments identifiés par des carrés tracés à la main dans les images, il a été nécessaire de *crowdsourcer* via Mechanical Turk les tâches vers des milliers d'annotateurs (Su et al., 2012 ; Jaton, 2017). De 9 298 à 14 millions de données, un tel bouleversement du volume des datasets – et donc des dimensions présentes dans les données – ne prend sens qu'accompagné par une croissance exponentielle de la puissance des calculateurs qui va être offert par la parallélisation et le développement des GPU (figure 6). En 2009, la « backprop » est implémentée sur des cartes graphiques permettant alors d'entraîner jusqu'à 70 fois plus vite un réseau de neurones (Raina *et al.*, 2009). Il est aujourd'hui considéré comme de bonne pratique d'apprendre une catégorie dans une tâche de classification avec 5 000 exemples par catégorie, ce qui conduit rapidement les jeux de données à avoir plusieurs millions d'exemples. La croissance exponentielle des jeux de données accompagne dans le même mouvement celle des architectures des calculateurs : le nombre de neurones dans un réseau double tous les 2,4 ans (Goodfellow *et al.*, 2016, p. 27).

Mais une autre transformation des données va aussi être mise en œuvre par les connexionnistes pour les granulariser et les traduire dans un format calculable en procédant à des opérations dites de « plongement » – ou « embedding ». Un réseau de neurones nécessite que les entrées du calculateur prennent la forme d'un vecteur. Le monde doit donc être préalablement codé sous la forme d'une représentation vectorielle purement numérique. Si certains objets tels que les images se décomposent naturellement en vecteurs, d'autres objets nécessitent d'être « plongés » dans un espace vectoriel avant d'être susceptibles d'être calculés ou classifiés par les réseaux de neurones. Il en va ainsi du texte qui constitue l'exemple prototype. Pour faire entrer un mot dans un réseau de neurones, la technique Word2vec le « plonge » dans un espace vectoriel qui mesure sa distance avec tous les autres mots du corpus (Mikolov *et al.*, 2013). Les mots héritent ainsi d'une position dans un espace de plusieurs centaines de dimensions. L'avantage d'une telle représentation réside dans les nombreuses opérations offertes par une telle transformation. Deux termes dont les positions inférées dans cet espace sont proches sont également similaires sémantiquement, on dit de ces représentations qu'elles sont distribuées : le vecteur du concept « appartement » [-0.2,

Figure 6. Croissance du nombre de données dans les datasets de recherche de 1990 à 2015 (en haut) et du nombre de neurones dans les architectures de calcul mises en place de 1960 à 2015.



Ces données ont été partiellement extraites de Goodfellow et al. (2016, pp. 21 et 24) et complétées par l'article de Wikipédia « List of dataset for Machine Learning »

0.3, -4.2, 5.1...] sera proche de celui du « maison » [-0.2, 0.3, -4.0, 5.1...]. La proximité sémantique n'est pas déduite d'une catégorisation symbolique, mais induite des voisinages statistiques entre tous les termes du corpus. Dès lors, ces vecteurs peuvent remplacer avantageusement les mots qu'ils représentent pour résoudre des tâches complexes comme la classification automatique de documents, la traduction ou le résumé automatique. Les concepteurs des machines connexionnistes procèdent donc bien à des opérations très artificielles pour traduire les données dans un autre système de représentation et les « brutifier » (Denis et Goëta, 2017). Si l'analyse automatique de la langue a été pionnière pour « plonger » des mots dans un espace vectoriel, on assiste aujourd'hui à une généralisation de la procédure de plongement (*embedding*) qui s'étend progressivement à tous les domaines applicatifs : les réseaux deviennent de simples points dans un espace vectoriel avec graph2vec, les textes avec paragraph2vec, les films avec movie2vec, le sens de mots avec sens2vec, les structures moléculaires avec mol2vec, etc. Selon la formule de Yann LeCun, l'ambition des concepteurs des machines connexionnistes est de mettre le monde dans un vecteur (*world2vec*). Au lieu de transformer les entrées en symboles articulés par un tissu de concepts interdépendants, cette vectorisation fabrique des voisinages entre des propriétés internes aux éléments du corpus d'apprentissage[31].

[31] Fidèles au modèle cognitif du connexionnisme, les trois principaux promoteurs du deep learning, Y. LeCun, G. Hinton et Y. Bengio en donnent la traduction calculatoire : « La question de la représentation est au cœur du débat entre les paradigmes de

Du modèle à l'architecture

Par un véritable mouvement de bascule, ce qui désormais est offert par la variété et le volume des données doit dès lors être soustrait du calculateur. Les concepteurs des architectures à neurones vont ainsi procéder à un évidement systématique et résolu de toutes règles explicites insérées « intentionnellement » dans les calculateurs pour identifier, caractériser ou agréger préalablement les données.

> « Il y a une force derrière ça, *explique un chercheur du domaine*. Il y a une vague qui est la vague des données, une espèce de grande vague de fond qui a tout emporté. Et cela a bazardé complètement tous les courants de pensée qui étaient basés sur de la modélisation humaine, sur de la modélisation explicite. J'ai travaillé sur plusieurs domaines qui sont des domaines applicatifs, de la parole, de l'écrit, du texte, des données sociales, et chaque fois j'ai vu la même chose. Les gens ont imaginé pendant une période mettre des connaissances dans leur système et cela a été balayé. Systématiquement ! Cela fait depuis trente ans que cela tombe, domaine par domaine. C'est comme ça. C'est un drôle de truc, tu vois. C'est la même chose que les mecs qui ont passé leur vie à croire dans un régime socialiste et puis ça s'effondre sous leurs pieds... C'est un truc du même ordre[32]. »

À partir de la fin des années 2000, le sentiment déstabilisant de voir une technique sans théorie se substituer aux efforts de modélisation patiemment conduits depuis des années va successivement traverser les communautés du signal, de la voix, de l'image et de la traduction automatique. Domaine après domaine, les calculs des réseaux de neurones gagnent en efficacité en transférant à la distribution des poids dans le réseau des opérations qui constituaient précédemment le principal foyer d'attention de l'activité scientifique : l'identification des caractéristiques (*feature engineering*) et la reconnaissance de formes (*pattern recognition*). Ces techniques consistent à programmer de façon « manuelle » (*handcrafted*) des algorithmes permettant d'identifier des caractéristiques des données initiales. Ce processus d'extraction facilite l'apprentissage en simplifiant la relation entre les caractéristiques et l'objectif du problème. L'automatisation de plus en plus forte de la découverte des caractéristiques va permettre aux techniques de *machine learning* statistiques de prendre le pouvoir sur les modélisateurs à l'intérieur des calculateurs (*supra*)[33]. Mais les réseaux de neurones radicalisent ce mouvement en écartant cette fois tout processus d'extraction de caractéristiques au profit d'un traitement appelé *end-to-end* (de bout en bout) : aller de la donnée numérique « brute » à l'exemple « étiqueté » sans faire de place à une intervention visant, de façon explicite, à produire des représentations intermédiaires des données guidant les calculs vers l'objectif.

---

la cognition inspirés par la logique et ceux inspirés par les réseaux de neurones. Dans le paradigme inspiré par la logique, une instance d'un symbole est quelque chose dont la seule propriété est qu'elle est identique ou non aux autres instances de symboles. Il n'a pas de structure interne pertinente pour son utilisation ; et pour raisonner avec des symboles, ils doivent être liés aux variables dans des règles d'inférence judicieusement choisies. En revanche, les réseaux de neurones n'utilisent que de grands vecteurs d'activité, de grandes matrices de poids et des non-linéarités scalaires pour effectuer le type d'inférence "intuitive" rapide qui sous-tend un raisonnement sans effort et de bon sens » (LeCun et al., 2015, p. 436).

[32] Interview F., un des pionniers du deep learning en France, 20 juillet 2018.

[33] « De nombreux développeurs de systèmes d'IA reconnaissent maintenant que, pour de nombreuses applications, il peut être beaucoup plus facile d'entraîner un système en lui montrant des exemples de comportements d'entrée-sortie désirés plutôt que de le programmer manuellement en anticipant la réponse désirée pour toutes les entrées possibles » (Jordan et Mitchell, 2015, p. 255).

Un exemple de ce déplacement est le principe de la convolution utilisé dans la vignette d'ouverture de cet article. La communauté de *computer vision* a développé de très subtiles méthodes d'extraction pour identifier au sein des images les bords, les coins, les transitions de contraste et des points d'intérêts particuliers dans les images, pour les associer à des sacs de mots servant de caractéristiques à la tâche confiée au calculateur. Ces opérations sont désormais prises en charge implicitement par la structure particulière donnée aux réseaux convolutifs : paver l'image en petites tuiles de pixels confiées à des segments de neurones séparés avant de les rassembler dans une autre couche du réseau. Plutôt que de modéliser un rhinocéros, ou les caractéristiques de blocs de pixels qui prédisent la forme rhinocéros, quelques milliers de photos de rhinocéros se déplaçant dans l'image, dont une partie du corps est tronquée, pris sous des angles et dans des positions variées, imprimeront beaucoup mieux la forme-concept « rhinocéros » dans les poids des neurones qu'une procédure de prétraitement des caractéristiques qui ne sait pas gérer les problèmes d'invariance d'échelle, de translation ou de rotation. La relation de la donnée à sa caractéristique n'est pas recherchée, mais obtenue. Le réseau de neurones procède bien à une extraction de caractéristiques, les bords sont souvent « vus » par la première couche de neurones, les coins par une autre, les éléments de formes plus complexes par une dernière, mais ces opérations, sans avoir été explicitement implémentées, sont des effets émergents du réseau sous contrainte d'architecture.

Le pré-traitement des « paramètres » du calcul a ainsi été transféré vers la définition des « hyper-paramètres » du calculateur. Plus la part de modélisation humaine décroît, plus devient complexe la spécification de l'architecture des machines inductives. Un réseau de neurones complètement connecté ne produit rien. Il est donc nécessaire de le sculpter afin d'adapter son architecture à la tâche d'apprentissage qui lui est confiée : nombre de couches cachées, de neurones par couche, plan des connexions, choix de la fonction d'activation, du type d'optimisation, des coefficients au début de l'apprentissage, choix d'une fonction d'objectif, nombre de fois où l'ensemble des données d'apprentissage sont montrées au modèle, etc. Ces réglages font l'objet d'ajustements par essais/erreurs. La technique d'élagage (*pruning*), par exemple, consiste à enlever des neurones pour voir si cela modifie la performance du réseau, celle du *dropout* suggère, lors de la phase d'apprentissage, de ne pas envoyer de signal vers certains neurones des couches d'entrée ou des couches cachées de façon aléatoire afin d'éviter le sur-apprentissage (*overfitting*) lorsque le réseau doit généraliser vers des données fraîches. Ces recettes, bonnes pratiques et règles de l'art nourrissent une bonne partie des discussions de la communauté et conservent un caractère artisanal (Domingos, 2012). Face aux raffinements mathématiques de l'extraction des *features*, la fabrication des réseaux de neurones peut ainsi apparaître comme un travail de hacker, une activité de programmeurs doués pratiquant avec habileté une sorte de magie noire.

> « Le truc qu'ils ont fait d'enlever toute l'extraction des caractéristiques pour prendre l'image brute, les mecs qui ont fait ça avec Hinton, c'est des dingues parce que c'est une chose de reproduire, mais d'y aller comme ça en explorant ! Ils ont fait des systèmes d'une complexité qu'on n'imaginait pas et ils ont été capables de les faire marcher. Tu prends un papier de ces gens, tu regardes,

> moi je suis effrayé, je suis trop vieux ! Les gars, ils te parlent comme s'ils programmaient presque. Ils ne font pas une description avec trois équations qui ont du sens pour moi. Mais en 5 lignes ils vont te décrire un machin qui est hypercomplexe. Donc, ça veut dire qu'il a fait une architecture dans laquelle il a mis 100 éléments les uns avec les autres et pour chacun, pour les relier, tu as dix choix possibles. Il a joué avec ça pour réussir à le faire tourner. C'est un hacker, c'est un boulot de hacker[34] ! »

Les hyper-paramètres sont donc le lieu où se sont déplacées les nouvelles exigences d'explicabilité des réseaux de neurones. Les données ne « parlent toutes seules » que soumises à une architecture qui, elle, ne peut être apprise des données et concentre désormais une grande part de la recherche en IA. À la conférence NIPS, un papier remarqué est un article qui propose une nouvelle architecture, à laquelle, comme pour les planètes, les chercheurs donnent systématiquement des noms composant ainsi un curieux bestiaire (figure 7). En déplaçant du modèle à l'architecture, le lieu où s'exprime l'inventivité des chercheurs, ce sont aussi les compétences et les qualités que requiert leur conception qui se transforment en permettant, notamment en raison de la disponibilité d'outils ouverts et faciles à manipuler, à une nouvelle population de *datascientists*, de bidouilleurs et de programmeurs d'entrer dans le champ précédemment très fermé des producteurs d'IA. En transformant l'architecture des machines prédictives, les connexionnistes ont ainsi contribué à déplacer les mondes sociaux de l'IA : d'abord, parce que les données « réelles », notamment celles venues des industries du numérique, se sont (partiellement) substituées aux *dataset* «jouets» des laboratoires académiques, ensuite parce que les savoir-faire requis pour fabriquer les machines connexionnistes appellent des compétences en développement informatique qui n'étaient pas celles des précédentes générations de l'IA.

LE TRAVAIL DE L'INDUCTION

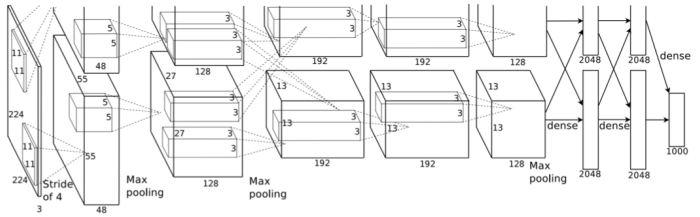
La trajectoire des machines intelligentes dont on vient de résumer l'histoire en quatre configurations successives témoigne de la profonde transformation de leur architecture (tableau 1 ci-dessous). Le *monde*, le *calculateur* et l'*horizon* de ces dispositifs ont été très profondément remaniés et les articulations entre ces composantes façonnent des dispositifs qui proposent des définitions sensiblement différentes de l'intelligence, du raisonnement et de la prédiction.

Tableau 1. Les quatre âges des machines prédictives

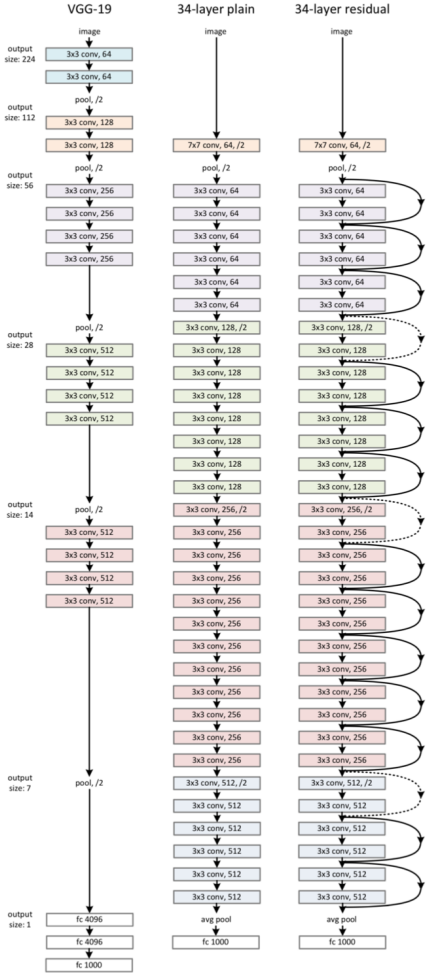
Machine	Monde	Calculateur	Horizon
Cybernétique (connexionniste)	*Environnement*	*« Boîte noire »*	*Negative feedback*
IA Symbolique (symbolique)	*Monde « jouet »*	*Raisonnement logique*	*Résolution de problème*
Système expert (symbolique)	*Monde de connaissances expertes*	*Sélection des hypothèses*	*Exemples/ contre-exemples*
Deep Learning (connexionniste)	*Le monde comme vecteur de données massives*	*Réseau de neurones profond*	*Optimisation de l'erreur sur objectif*

[34] Interview F., un des pionniers du deep learning en France, 20 juillet 2018.

Figure 7. Exemples de trois architectures de réseaux de neurones victorieuses du challenge ILSVRC de 2012 à 2015



(a) 9 couches du réseau AlexNet (ImageNet 2012)



(b) 40 couches de GoogLeNet (ImageNet 2014)

(c) 152 couches par Microsoft (ImageNet 2015)

Une dynamique d'ensemble apparaît cependant dans cette histoire mouvementée. Le projet matérialiste d'une fabrication computationnelle de l'esprit s'est aujourd'hui engagé dans une voie résolument connexionniste. Les succès actuels des machines inductives ne signifient en rien qu'un terme ou une « solution » ait été trouvé(e). En dépit de leurs prouesses, les techniques de *deep learning* sont très loin de satisfaire le programme de l'intelligence artificielle générale, comme ne cessent de leur reprocher les « symbolistes »

réclamant, la main agrippée au bord de la falaise, une hybridation entre les approches[35]. Mais ce qui apparaît surtout dans la trajectoire mise en récit dans cet article est que cette recomposition inductive du calcul prédictif ne pouvait se réaliser sans que soit entrepris un considérable et ambitieux travail visant à modifier l'équilibre entre le monde des données et la forme du calcul.

En entrée du calculateur, d'abord, la composition du monde a subi un profond mouvement d'atomisation et de granularisation. Alors que les mondes « jouets » et les mondes de connaissances des *machines symboliques* étaient faits de petits univers restreints, nettoyés et domestiqués par une grille de caractéristiques intelligibles et interdépendantes, les *machines connexionnistes* se déploient dans un monde dans lequel les données doivent non seulement être massives, mais aussi les plus atomisées possible afin de leur ôter toute structure explicite. Si les données enferment bien des régularités, des relations compositionnelles, des styles globaux, etc., ceux-ci doivent être mis en évidence par le calculateur et non par le programmeur. Le premier trait du travail de production de l'induction consiste donc à faire entrer dans le système des données sous la forme la plus élémentaire possible : des pixels plutôt que des formes, des fréquences plutôt que des phonèmes, des lettres plutôt que des mots, des clics plutôt que des déclarations d'internautes, des comportements plutôt que des catégories... (Cardon, 2017). Que les données soient hétérogènes, redondantes et souvent incorrectes n'apparaît plus comme un problème, chaque nouveau signal peut être ajouté sous la forme d'une nouvelle colonne dans la matrice d'entrée qui forme le monde des machines connexionnistes. Ce n'est donc pas une nature « brute » et « immédiate » qui se rend disponible à la perception des calculateurs, mais le produit d'un travail d'atomisation et de désassociation des données afin d'en faire des signes numériques normalisés les plus élémentaires possible. Pour fabriquer ces entrées, une nouvelle métrologie de capteurs, d'enregistrements et de bases de données constitue une infrastructure indispensable à la traduction des images, sons, déplacements, clics ou variables de toutes sortes en ces gigantesques vecteurs dont ont besoin les machines connexionnistes (Mackenzie, 2017).

La deuxième caractéristique de ce mouvement d'ensemble est la disparition d'une modélisation *a priori* des activités du calculateur (phénomène souvent décrit comme « fin de la théorie » (Anderson, 2008)) au profit d'abord d'une probabilisation des modèles au sein d'un espace d'hypothèses de plus en plus large, puis d'une dispersion plus radicale des modèles lorsque la prise en compte des dimensions variées des données se trouve distribuée au sein des multiples couches des réseaux de neurones. C'est l'immense ambition intellectuelle déployée par l'IA des premiers âges pour modéliser le raisonnement qui a sombré tout en laissant au passage d'importantes contributions à la recherche en informatique. Les machines connexionnistes ont déplacé les enjeux de l'IA de la résolution de problèmes abstraits, objets des sciences cognitives orthodoxes, à la perception de caractéristiques au sein d'énormes masses de signaux sensibles. Le deuxième trait du travail

[35] Voir le débat entre Y. LeCun et G. Markus (2017). Ce dernier en appelle à une hybridation des approches symbolique et connexionniste car cette dernière présente de nombreuses faiblesses qui dessinent les nouveaux enjeux de recherche du domaine. Elle permet d'interpoler entre eux deux exemples connus, mais elle extrapole mal dans des situations qui n'ont pas fait l'objet d'apprentissage ; ses modèles consomment un nombre considérable de données étiquetées qui sont loin d'être toujours accessibles ; elle ne sait pas hiérarchiser un raisonnement en isolant des règles et des abstractions ; elle n'est pas en mesure d'intégrer un savoir préexistant relatif aux données calculées ; elle manque de transparence et d'explicabilité ; elle prédit dans un monde stable et statique sans être préparé à l'imprévu ; elle est probabiliste et ne sait prédire avec certitude et précision (Markus, 2018).

de production de l'induction aura ainsi été de parvenir à réunir les conditions permettant de renverser le dispositif calculatoire de l'IA afin de faire des programmes des sorties et non des entrées. Pour autant, les réseaux de neurones ne font aucunement disparaître la « théorie ». Ils ne font que la déplacer vers les hyper-paramètres de l'architecture du calculateur, tout en donnant à l'idée de « théorie » une signification sans doute moins « symbolisable ». Cette question rend particulièrement délicats les enjeux relatifs à la compréhension et à l'interprétabilité du processus qu'ils mettent en œuvre pour effectuer leur prédiction (Burrel, 2016 ; Cardon, 2015). Comme y invitaient le PDP dans les années 1980 et beaucoup des travaux sur les systèmes complexes, sans doute nous faut-il apprendre à rendre perceptibles, appropriables et discutables des formes de modélisation qui n'ont plus les propriétés (linéarité, lisibilité, complétude, parcimonie, etc.) auxquelles nous a habitués l'idée – très « symbolique » – d'intelligibilité des modèles dans les sciences sociales.

Le troisième déplacement a trait à l'horizon donné au calculateur. Alors que la machine intelligente imaginée par l'IA symbolique se donnait comme objectif les attendus rationnels de la logique – rationalité interne aux calculs qui a permis aux promoteurs de l'IA de soutenir que les machines étaient « autonomes » –, dans le modèle connexionniste l'horizon du calcul n'appartient pas au calculateur, mais au monde qui lui a donné des exemples « étiquetés ». Les sorties – produites, symbolisées et biaisées par les humains – constituent aujourd'hui l'une des plus précieuses entrées des machines connexionnistes. Le troisième trait du travail de production de l'induction est d'asseoir la performance de la prédiction sur le monde lui-même en renouvelant la promesse adaptative des *machines reflets* de la cybernétique : faire système avec l'environnement à calculer pour installer des boucles de rétroaction d'un nouveau type. Il y a donc quelque paradoxe à constater que, perpétuant une conception « symbolique » de l'intelligence des machines, une grande partie des débats critiques sur les biais des nouvelles formes de calcul se porte vers les intentions stratégiques des programmateurs alors même que ceux-ci ne cessent de chercher à effacer toute trace d'introjections « humaines » (*knowledge free*) préalables dans les opérations du calculateur. Il est certes judicieux d'être très vigilant à l'égard des objectifs stratégiques que les entreprises de l'économie numérique cherchent à glisser dans les calculs réalisés par leurs services. Cependant, pour être plus pertinente et efficace, sans doute serait-il nécessaire que la critique adapte aussi ses prises à la révolution « inductive » des machines à prédire. Car si les prédictions calculées ne sont pas le reflet « naturel » des données, c'est de plus en plus à la composition des données en entrée, à l'architecture retenue par les différents systèmes et aux objectifs qui ont été donnés à la supervision de l'apprentissage qu'il est nécessaire de prêter attention. Apologétiques ou critiques, les représentations de l'Intelligence artificielle qui ont été nourries par une science-fiction puisant son imaginaire dans l'IA symbolique – Marvin Minsky était le conseiller scientifique de *2001, l'Odyssée de l'espace* ! – apparaissent grandement désajustées, désuètes et, somme toute, peu imaginatives face à la réalité beaucoup plus intrigante et originale de ces nouvelles machines.

RÉFÉRENCES

ANDERSON C. (2008), *The end of theory: Will the data deluge makes the scientific method obsolete?*, June 23, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

ANDERSON J. A., ROSENFELD E. (eds.) (1988), *Neurocomputing: Foundations of Research*, Cambridge, The MIT Press.

ANDLER D. (1990), « Connexionnisme et cognition. À la recherche des bonnes questions », *Revue de synthèse*, n° 1-2, pp. 95-127.

ANDLER D. (1992), « From paleo to neo-connectionism », in G. VAN DER VIJVER (ed.), *Perspectives on Cybernetics*, Dordrecht, Kluwer, pp. 125-146.

ANDLER D. (2016), *La silhouette de l'humain. Quelle place pour le naturalisme dans le monde d'aujourd'hui ?*, Paris, Gallimard.

ASHBY R. (1956), *Introduction to Cybernetics*, London, Chapman & Hall.

BENGIO Y., COURVILLE A., VINCENT P. (2013), « Representation Learning: A Review and New Perspectives », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, n° 8.

BENGIO Y., LECUN Y. (2007), « Scaling Learning Algorithms towards AI », in L. BOTTOU, O. CHAPELLE, D. DECOSTE, J. WESTON, *Large-Scale Kernel Machines*, Cambridge, MIT Press.

BOSER B. E., GUYON I. M., VAPNIK V. N. (1992), « A Training Algorithm for Optimal Margin Classifiers », *Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, pp. 144-152.

BREIMAN L. (2001), « Statistical Modeling: The Two Cultures », *Statistical Science*, vol. 16, n° 3, pp. 199-215.

BROOKS R. A. (1988), « Intelligence without Representation », *Mind Design*, in J. HAUGELAND (ed.), *Mind Design*, Cambridge MA, The MIT Press.

BURRELL J. (2016), « How the machine 'thinks': Understanding opacity in machine learning algorithms », *Big Data & Society*, January-June, pp. 1-12.

CARBONELL J. G., MICHALSKI R. S., MITCHELL T. (1983), « Machine Learning: A Historical and Methodological Analysis », *AI Magazine*, vol. 4, n° 3, pp. 69-79.

CARDON D. (2015), *À quoi rêvent les algorithmes. Promesses et limites*, Paris, Seuil, coll. « République des idées ».

CARDON D. (2017), « Infrastructures numériques et production d'environnements personnalisés », in K. CHATZISI, G. JEANNOT, V. NOVEMBER, P. UGHETTO (dir.), *Les métamorphoses des infrastructures, entre béton et numérique*, Bruxelles, Peter Lang, pp. 351-368.

COLLINS H. M. (1992), *Experts artificiels. Machines intelligentes et savoir social*, Paris, Seuil.

CORNUÉJOLS A., MICLET L., BARRA V. (2018), *Apprentissage artificiel. Concept et algorithmes*, Paris, Eyrolles (3e éd.).

CREVIER D. (1997), *À la recherche de l'intelligence artificielle*, Paris, Champs/ Flammarion [1re éd. américaine 1993].

DAGIRAL É., PARASIE S. (2017), « La "science des données" à la conquête des mondes sociaux. Ce que le "Big Data" doit aux épistémologies locales », in P.-M. MENGER, S. PAYE (dir.), *Big data et traçabilité numérique. Les sciences sociales face à la quantification massive des individus*, Paris, Collège de France.

DENG J., DONG W., SOCHER R., LI L. J., LI K., FEI-FEI L. (2009). « Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition », *CVPR 2009*, pp. 248-255.

DENIS J., GOËTA S. (2017), « Les facettes de l'Open Data : émergence, fondements et travail en coulisses », in P.-M. MENGER, S. PAYE (dir.), *Big data et traçabilité numérique. Les sciences sociales face à la quantification massive des individus*, Paris, Collège de France.

DOMINGOS P. (2012), « A Few Useful Things to Know about Machine Learning », *Communication of the ACM*, vol. 55, n° 10, pp. 78-87.

DOMINGOS P. (2015), *The Master Algorithm. How the question for the ultimate machine will remake our world*, London, Penguin Random House UK.

DREYFUS H. (1972), *What Computers Can't Do: The Limits of Artificial Intelligence*, New York, Harper and Row.

DREYFUS H. (2007), « Why Heideggerian AI failed and how fixing it would require making it more Heideggerian », *Artificial Intelligence*, n° 171, pp. 1137-1160.

DUPUY J.-P. (2005), *Aux origines des sciences cognitives*, Paris, La Découverte.

EDWARDS P. N. (1996), *The Closed World. Computers and the Politics of Discourses in Cold War America*, Cambridge MA, The MIT Press.

FEIGENBAUM E. A., McCORDUCK P. (1983), *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*, Reading, Addison Wesley.

FLECK J. (1982), « Development and Establishment in Artificial Intelligence », in N.ELIAS, H.MARTINS, R.WHITLEY (eds.), *Scientific Establishments and Hiérarchies, Sociology of the Sciences Yearbook*, vol. 6, Dordrecht, Reidel, pp. 169-217.

FLECK J. (1987), « Postscript: The Commercialisation of Artificial Intelligence », in B. BLOMFIELD (ed.), *The Question of AI*, London, Croom-Helm, pp. 149-64.

FODOR J. A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*, Cambridge MA, MIT Press.

GALISON P. (1994), « The ontology of the enemy: Norbert Wiener and the cybernetic vision », *Critical Inquiry*, vol. 21, n° 1, pp. 228-266.

GARDNER H. (1985), *The Mind's New Science. A History of Cognitive Revolution*, New York, Basic Books.

GITELMAN L. (ed.) (2013), *Raw data is an oxymoron*, Cambridge MA, MIT Press.

GOLDSTEIN I., PAPERT S. (1977), « Artificial Intelligence. Language and the Study of Knowledge », *Cognitive Science*, vol. 1, n° 1.

GOLDSTINE H. (1972), *The Computer From Pascal to Von Neumann*, Princeton, Princeton University Press.

GOODFELLOW I., BENGIO Y., COURVILLE A. (2016), *Deep Learning*, Cambridge MA, MIT Press.

GRUDIN J. (2009), « AI and HCI: Two fields divided by a common focus », *AI Magazine*, vol. 30, n° 4, pp. 48-57.

HAUGELAND J. (1985), *Artificial Intelligence: The Very Idea*, Cambridge MA, MIT Press.

HEBB D. O. (1949), *The Organization of Behavior*, New York, Wiley.

HOPFIELD J. J. (1982), « Neural Networks and Physical Systems with Emergent Collective Computational Abilities », *Proc. Natl. Acad. Sc. USA*, vol. 79.

JATON F. (2017), « We get the algorithms of our ground truths: Designing referential databases in Digital Image Processing », *Social Studies of Science*, vol. 47, n° 6, pp. 811-840.

JORDAN M. (2018), « Artificial Intelligence: The Revolution hasn't happened yet », *Medium*, April 19.

JORDAN M. I, MITCHELL T. M. (2015), « Machine learning: Trends, perspectives, and prospects », *Science*, vol. 349, n° 6245, pp. 255-260.

KRIZHEVSKY A., SUTSKEVER I., HINTON G. (2012), « ImageNet Classification with Deep Convolutional Neural Networks », *NIPS 2012*, Lake Tahoe, December 3-6.

KURENKOV A. (2015), « A 'Brief' History of Neural Nets and Deep Learning », *andreykurenkov.com*, December 24.

LATOUR B. (1987), *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge MA, Harvard University Press.

LECUN Y. (2007), « Who is Afraid of Non-Convex Loss Functions? », *2007 NIPS workshop on Efficient Learning*, Vancouver, December 7.

LECUN Y., BENGIO Y., HINTON G. (2015), « Deep learning », *Nature*, vol. 521, n° 7553.

LECUN Y., BOSER B., DENKER J., HENDERSON D., HOWARD R., HUBBARD W. JACKEL L. (1989), « Backpropagation Applied to Handwritten Zip Code Recognition », *Neural Computation*, vol. 1, n° 4, pp. 541-551.

LECUN Y., MARKUS G. (2017), «Debate: "Does AI Need More Innate Machinery?" », *YouTube*, October 20.

MARKOFF J. (2015), *Machines of loving grace. Between human and robots*, HarperCollins Publishers, 2015.

MACKENZIE A. (2017), *Machine Learners. Archaeology of a Data Practice*, Cambridge MA, The MIT Press.

MARKUS G. (2018), « Deep Learning: A Critical Appraisal », *arXiv :1801.00631*, January 2.

MAZIÈRES, A. (2016). *Cartographie de l'apprentissage artificiel et de ses algorithmes*. Manuscrit de thèse, Université Paris Diderot.

McCARTHY J. (1988), « [Review of] Bloomfield Brian ed. The Question of Artificial Intelligence... », *Annals of the History of Computing*, vol. 10, n° 3, pp. 221-233.

McCORDUCK P. (1979), *Machines Who Think. A Personal Inquiry into the History and Prospects of Artificial Intelligence*, Natick, AK Peters.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S., DEAN J. (2013), « Distributed representations of words and phrases and their compositionality », *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 3111-3119.

MILLER G. A. (1995), « WordNet: A Lexical Database for English », *Communications of the ACM*, vol. 38, n° 11, pp. 39-41.

MINSKY M. (1975), « A Framework for Representing Knowledge », in P. WINSTON (ed.), *The Psychology of Computer Vision*, New York, McGraw-Hill.

MINSKY M. (1986), *The Society of Mind*, New York, Simon & Schuster.

MINSKY M., PAPERT S. (1969), *Perceptrons: An Introduction to Computational Geometry*, Cambridge MA, The MIT Press.

MINSKY M., PAPERT S. (1970), « Draft of a Proposal to ARPA for Research on Artificial Intelligence at MIT, 1970-1971 », *Artificial Intelligence Lab Publication*, MIT.

MITCHELL T. (1977), « Version Spaces: A Candidate Elimination Approach to Rule Learning », *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, Cambridge, August, pp. 305-310.

NEWELL A., SIMON H., SHAW J. C. (1956), « The Logic Theory Machine », *IRE Transactions on Information Theory*, vol. IT-2, n° 3.

NEWELL A., SIMON H. A. (1963), « GPS: A Program That Simulates Human Thought », in E. A. FEIGENBAUM, J. FELDMAN (eds.), *Computers and Thought*, New York, McGraw-Hill, pp. 279-283.

NILSSON N. J. (2010), *The Quest for Artificial Intelligence. A history of ideas and achievements*, Cambridge, Cambridge University Press.

OLAZARAN M. (1996), « A Sociological Study of the Official History of the Perceptron Controversy », *Social Studies of Science*, vol. 26, n° 3, pp. 611-659.

PICKERING A. (2010), *The Cybernetic Brain. Sketches of another Future*, Chicago IL, The Chicago University Press.

RAINA R., MADHAVAN A., NG A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, ACM, pp. 873-880.

RID T. (2016), *Rises of the Machines. The lost history of cybernetics*, London, Scribe Publications.

ROLAND A., SHIMAN P. (2002), *Strategic Computing. DARPA and the Quest for Machine Intelligence, 1893-1993*, London, The MIT Press.

ROSENBLUETH A., WIENER N., BIGELOW J., (1943), « Behavior, Purpose and Teleology », *Philosophy of Science*, vol. 10, n° 1, pp. 18-24.

RUMELHART D. E., HINTON G., WILLIAMS R. J. (1986a), « Learning representations by back-propagating errors », *Nature*, n° 323, pp. 533-536.

RUMELHART D. E., McCLELLAND J. L. (1986b), « PDP Models and General Issues in Cognitive Science », in PDP RESEARCH GROUP (1986), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, Cambridge MA, MIT Press.

SHANNON C. (1948), « A mathematical theory of communication », *Bell System Technical Journal*, n° 27, pp. 379-423.

SKINNER B. F. (1971), *Beyond Freedom and Dignity*, New York, Bantam.

SMOLENSKY P. (1988), « The proper treatment of connectionism », *The Behavioral and Brain Sciences*, vol. 11, pp. 1-74.

SU H., DENG J., FEI-FEI L. (2012), « Crowdsourcing Annotation for Visual Object Detection », *AAAI Workshops*, Toronto.

TRICLOT M. (2008), *Le moment cybernétique. La constitution de la notion d'information*, Paris, Champ Vallon.

TUKEY J. W. (1962), « The future of data analysis », *The Annals of Mathematical Statistics*, vol. 33, n° 1, pp. 1-67.

TURING A. (2004), « Proposal for Development in the Mathematics of an Automatic Computing Engine (ACE) », in J. COPELAND (ed.), *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and And Artificial Life plus The Secret of Enigma*, New York, Oxford University Press.

VON NEUMANN J. (1945), « First Draft of a Report on EDVAC », *Contract n° W-670-ORD-4926 Between the United States Army Ordnance Department and the University of Pennsylvania*, Moore School of Electrical Engineering.

WALDROP M. (1987), *Man-made Minds: The Promise of Artificial Intelligence*, New York, Walker.

WEIZENBAUM J. (1976), *Computer and Human Reason*, San Francisco, Freeman.

WIENER N. (1948), *Cybernetics, or control and communication in the animal and the machine*, Cambridge, Cambridge University Press.

WINOGRAD T. (1972), *Understanding Natural Language*, Edinburgh, Edinburgh University Press.

WINOGRAD T., FLORES F. (1986), *Understanding Computers and Cognition: A New Foundation for Design*, Norwood, Ablex Publishing Corporation.