Intelligence artificielle et intuition

Alban Leveau-Vallier

UNIVERSITÉ DE PARIS VIII VINCENNES À SAINT-DENIS

LABORATOIRE D'ÉTUDES ET DE RECHERCHES SUR LES LOGIQUES CONTEMPORAINES DE LA PHILOSOPHIE (LLCP – EA4008)

ÉCOLE DOCTORALE PRATIQUE ET THÉORIE DU SENS (ED 31)

INTELLIGENCE ARTIFICIELLE ET INTUITION

LES ALGORITHMES D'APPRENTISSAGE PROFOND COMME OCCASION DE DÉCRIRE L'INTUITION

THÈSE

PRÉSENTÉE ET SOUTENUE PUBLIQUEMENT PAR

Alban LEVEAU-VALLIER

LE 27 JANVIER 2023

Pour l'obtention du **DOCTORAT EN PHILOSOPHIE**

Sous la direction de Pierre CASSOU-NOGUÈS

JURY

David BATES	Professeur à l'Université de Californie Berkeley
Pierre CASSOU-NOGUÈS	Professeur à l'Université Paris VIII
Béatrice JOYEUX-PRUNEL	Professeure ordinaire à l'Université de Genève
Jean LASSÈGUE	Directeur de recherche au CNRS
Giuseppe LONGO	Directeur de recherche au CNRS
François SEBBAH	Professeur à l'Université Paris X

SOMMAIRE

REMERCIEMENTS

La stimulante lecture des travaux de Pierre Cassou-Noguès sur Gödel a été un des éléments déclencheurs de mon intérêt pour la question de la réductibilité de l'esprit à une machine. Je lui sais gré d'avoir spontanément accepté de diriger cette thèse et de m'avoir accompagné de son écoute patiente, ses conseils avisés et son pragmatisme.

Je remercie David Bates – qui a eu l'amabilité de prendre le temps d'échanger autour de ses travaux –, Béatrice Joyeux-Prunel – dont les cours à l'ENS ont été un des moments forts de ma scolarité et dont le cycle sur l'imagination artificielle m'a aidé à mieux percevoir les liens entre mes recherches et les questions artistiques, sans rester cantonné à la philosophie –, Jean Lassègue – dont les ouvrages m'ont permis d'élargir mon interprétation des textes de Turing –, Giuseppe Longo – dont les travaux sur la notion de loi de la nature ont été l'occasion d'intuitions déterminantes –, ainsi que François Sebbah – que j'ai eu la chance d'entendre et de rencontrer à Cerisy, puis au Cube –, pour avoir eu la générosité d'accepter de participer à l'évaluation de ce travail qui leur doit beaucoup.

Mes remerciements vont également aux amis informaticiens et spécialistes de l'intelligence artificielle qui ont bien voulu prendre le temps de m'expliquer en quoi consiste leur pratique et de répondre à mes questions : Gautier Boucher (notamment lors de nos années de « coworking » à l'atelier d'Aligre), Julien Murésianu, Laurent Sifre, et en particulier Karl Neuberger et Julien Vong de Quantmetry, qui, grâce aux formations réalisées et dispensées ensemble, m'ont permis de mieux saisir comment s'utilise l'intelligence artificielle aujourd'hui. Je dois une mention spéciale au très sollicité Yann LeCun qui a eu la gentillesse de m'accorder un long entretien afin de détailler son concept d'intuition ainsi que sa participation à l'histoire du *deep learning*.

Mes réflexions ne seraient rien sans les nombreuses discussions avec les collègues et amis de disciplines variées : Gregory Chatonsky, dont le séminaire à Paris 8 a contribué à mettre en route mes premières idées, Daniel Andler, qui, au gré de nos différentes collaborations, a toujours chaleureusement répondu à mes questions et m'a aidé à recontextualiser les débats actuels dans le cadre plus large de l'histoire de l'intelligence artificielle, Guillaume Lamy pour ses retours attentifs et stimulants, et surtout Raphaëlle Brin pour ses patientes relectures et son

immense générosité. Je remercie celles et ceux qui m'ont offert l'opportunité de présenter et de discuter mes recherches lors de colloques, de conférences, ou de publications communes : Arnaud Regnauld, Monica Michlin, Hélène Machinal, Sylvie Bauer, Gwen Le Cor, Stéphane Vanderhaeghe, Emmanuel Basset, Karim Ghorbal, Yves Citton, Hélène Delahaye ; celles et ceux qui m'ont invité à exposer mes travaux dans leurs organisations respectives : Lucile Hofman, Norah Manasseh, Bibi Ndiaye, Bastien Guerry, Florian Marsaud, Loïc Brient, Gilbert Macquart. Je suis particulièrement reconnaissant à Pierre Cohen-Tanugi pour ses sollicitations dans le cadre de l'Institut de l'ENS qui m'ont permis d'approfondir mes connaissances sur l'intelligence artificielle, mais aussi de les enrichir par l'étude de nombreuses thématiques voisines. Je sais gré de la confiance qu'ils m'ont témoignée en me permettant d'enseigner dans leurs établissements : Marie Fidelin, Jimmy Mirande et Eric Buignet à l'ISCOM ; Sonia Jeanson et Dieudonné Abboud à l'ISEP ; Corinne Leforestier et Joffrey Lavigne à Sciences Po Paris.

Je salue mes compagnons d'infortune, doctorants ou post-doctorants, et les remercie pour leur écoute et leur soutien : Sara Touiza Ambroggiani, François Levin, Fabien Ferri, Joseph Lemelin ainsi que les membres des différents *reading groups* et groupes de doctorants auxquels j'ai eu la chance de participer.

Enfin, je suis reconnaissant à celles et ceux qui, tout au long de ces années, m'ont soutenu de près ou de loin par leur amitié et leur bienveillance, avec une mention spéciale pour ma famille, Sam et l'équipe des grandes occasions, celle des Grands Voisins, Rudy, Faustine, Axelle, Cécile, Olivier, Raphaëlle, Jérémie, Zoé, Sophie D'Aulan, Marie Boulet, Yael Azoulay. *Last but not least*, je n'ai pas de mots pour exprimer ma gratitude à Morgane pour son amour, sa patience et son soutien indéfectible ; ainsi qu'à Alice, dont la naissance, pendant la rédaction de ce travail, est venue bouleverser et illuminer mon existence.

RÉSUMÉ

Intelligence artificielle et intuition, les algorithmes d'apprentissage profond comme occasion de décrire l'intuition

Grâce aux succès techniques remportés par l'école connexionniste dans les années 2010, dont la victoire d'AlphaGo, la notion d'intuition, longtemps mise à l'écart par l'école symbolique, a repris droit de cité dans le champ de l'intelligence artificielle (IA). Selon leurs inventeurs, les algorithmes dits « d'apprentissage profond » (*deep learning*) font preuve d'intuition. Afin d'apprécier la pertinence de cette affirmation, cette thèse présente les réseaux de neurones d'apprentissage profond – leur histoire, leur fonctionnement, leur usage –, et s'arrête sur leur manière de simuler certaines de nos facultés (perception, induction, imagination), puis propose une définition du projet d'intelligence artificielle. En aspirant à matérialiser l'intelligence, et plus précisément l'intuition et la créativité, le projet d'intelligence artificielle s'immisce dans le champ de la philosophie et invite à rouvrir la question de la genèse de la pensée : suffit-il de réunir certaines conditions matérielles pour susciter l'intelligence, ou bien d'autres « ingrédients », dont l'intuition serait le nom, sont nécessaires ? En s'appuyant sur l'histoire de la philosophie (le différend entre Descartes et Leibniz autour de la « pensée aveugle »), sur des textes de penseurs plus contemporains (Bachelard, Bateson, Caillois, Derrida, Latour, Rosset), des fondateurs du projet d'IA (McCarthy, Turing) et de chercheurs actuels (Hinton, LeCun) cette thèse offre une description de l'intuition, de phénomènes récalcitrants à la simulation (compréhension, réflexivité, invention), et mène une réflexion sur le hasard afin de statuer sur la possibilité, pour les machines, d'être à l'origine de formes nouvelles.

Mots clés : intelligence artificielle, IA, apprentissage profond, intuition, créativité, hasard

SUMMARY

Artificial intelligence and intuition, deep learning as an occasion to describe intuition

Since the years 2010, and thanks to a number of feats achieved by connectionist artificial intelligence (like AlphaGo's victory), the notion of intuition, which had been put aside by the symbolic approach, has recovered a new legitimacy. According to their inventors, deep learning algorithms manifest intuition. In order to assess this claim, this thesis describes deep learning – its history, operating, and usage –, highlights how it simulates some human abilities (perception, induction, imagination) and provides a definition of artificial intelligence. As it pretends to materialize intelligence, and more specifically intuition and creativity, the project of artificial intelligence steps on the toes of philosophers and constrains them to reopen the question of mind genesis : would gathering the adequate material conditions be enough in order for intelligence to arise, or are there some missing « ingredients », of which intuition is the name ? Drawing on the history of philosophy, like the debate between Descartes and Leibniz about formalism, as well as discussions of contemporary thinkers (Bachelard, Bateson, Caillois, Derrida, Latour, Rosset), AI founders (McCarthy, Turing), and deep learning inventors (LeCun, Hinton), this thesis offers a description of intuition and of some features (comprehension, reflexivity, invention) that seem reluctant to simulation. Finally, it examines the notion of randomness in order to assess the possibility, for a machine, to originate new forms.

Keywords : artificial intelligence, AI, deep learning, intuition, creativity, randomness

INTRODUCTION

« Je crois que l'intuition humaine est encore en avance sur l'IA et je ferai de mon mieux pour maintenir l'avancée de l'intelligence humaine[1]. » En présentant ainsi son match contre AlphaGo, Lee Sedol relaye l'idée que l'intuition serait le privilège de l'humain et la part non machinique de l'intelligence, cet « oracle » dont Turing écrivait en 1938 qu'« il ne peut être une machine[2] », et donc une objection à la conjecture fondatrice formulée à l'occasion du séminaire de Dartmouth selon laquelle tous les aspects de l'intelligence peuvent être simulés[3]. Toutefois, pour que l'intuition joue le rôle du « résultat négatif » invalidant les prétentions du projet d'intelligence artificielle (ou IA) aussi sûrement que la seconde loi de la thermodynamique a pu invalider les ambitions du mouvement perpétuel, il faudrait en donner une description claire. Or, à commencer par Turing, ceux qui font référence à l'intuition se dispensent généralement de la décrire autrement que par la négative[4] et d'en donner d'autres caractéristiques que sa différence d'avec la machine et notre incapacité à la décrire, prêtant ainsi le flanc à la critique d'un refus par principe, que Catherine Malabou désigne ironiquement comme la « tortue » des philosophes, en référence à la célèbre posture défensive de l'armée romaine[5].

Le match entre AlphaGo et Lee Sedol est très largement couvert par la presse, qui rabâche un discours bien rodé : il y a plus de possibilités au go que d'atomes dans l'univers[6]. Jouer au go relève de l'incalculable, donc de l'intuition. Une victoire d'AlphaGo signifierait que les machines sont devenues intuitives, que plus rien ne les différencie des humains. Demis

1 Greg Kohs, *AlphaGo*, Netflix, 2018, 90 minutes.

2 « We shall not go any further into the nature of this oracle apart from saying that it cannot be a machine. » Alan Turing, « Systems of Logic Based on Ordinals », in Jack 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, 2004, p. 156.

3 « The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. » John McCarthy, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon, « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 », *AI Magazine*, vol. 27, No. 4, 2006.

4 « I shall not attempt to explain this idea of ''intuition'' any more explicitly. » Alan Turing, *op. cit.*, p. 192.

5 Catherine Malabou, *Métamorphoses de l'intelligence,* Paris, Presses Universitaires de France, 2017, p. 59.

6 « '*Les règles sont très simples, mais il s'agit probablement du jeu le plus complexe inventé par l'homme, car le nombre de combinaisons possibles est supérieur au nombre d'atomes dans l'univers […]',* estime le neuroscientifique anglais Demis Hassabis, cofondateur de DeepMind », Camille Gévaudan, « Le jeu de go pour les nuls », *Libération,* 8 mars 2016, https://www.liberation.fr/futurs/2016/03/08/le-jeu-de-go-pour-les-nuls_1438397/ page consultée le 20 novembre 2020.

Hassabis, l'entrepreneur à l'origine du programme, le répète à qui veut l'entendre : AlphaGo, basé sur des réseaux de neurones convolutifs, « imite ce que font les gens avec leur intuition[7] ». Suite à la victoire d'AlphaGo, Yann LeCun, inventeur des réseaux de neurones convolutionnels, déclare que le programme aurait fait preuve d'« une espèce d'intuition de ce qu'il faut jouer[8] ». Stanislas Dehaene, titulaire de la chaire de psychologie cognitive expérimentale au Collège de France, renchérit :

> le logiciel AlphaGo ne se contente pas d'une exploration arborescente mécanique, comme le faisait déjà *Deep Blue* pour les échecs il y a vingt ans. Il développe aussi une intuition visuelle des coups les plus pertinents. Pour moi, c'est éminemment proche de l'intuition humaine[9].

Alors que les échecs sont un jeu de méthodes explicites, le go requiert un sens stratégique plus diffus. Alors que Deep Blue fonctionnait par recherche arborescente, AlphaGo se base sur la reconnaissance de formes. En d'autres termes, les échecs sont le jeu paradigmatique de l'école symbolique (GOFAI) tandis que le go caractérise l'approche connexionniste. La victoire éclatante d'AlphaGo marque aussi le triomphe de cette dernière, longtemps mise au ban de la recherche en intelligence artificielle, et le retour de la notion d'intuition.

D'après les tenants de l'école symbolique, l'ensemble des processus cognitifs peut se décrire comme une manipulation de symboles selon des règles de syntaxe, ce qui revient à « identifier la pensée symbolique à la pensée » et « à nier l'existence d'une quelconque intuition[10] ». Les tenants de l'école connexionniste défendent l'opinion inverse : l'intuition existe, seulement son mécanisme n'est pas accessible à la conscience. Les solutions et les idées viennent à nous sans que nous sachions comment elles se forment car elles sont le résultat d'une multitude d'opérations neuronales inconscientes effectuées sur la base d'expériences sensorielles. En fabriquant des machines qui simulent la plupart des processus cognitifs (perception auditive et visuelle, conduite de véhicules, utilisation du langage, résolution de problèmes mathématiques…) comme étant le résultat d'une telle activité neuronale

7 « […] some kind of computer algorithm to mimic what people do with their intuition », Demis Hassabis dans Greg Kohs, *AlphaGo, op. cit.*

8 « Oui, en effet, les réseaux convolutifs. Ils ont donc utilisé ces méthodes récentes pour permettre aux machines d'avoir une espèce d'intuition de ce qu'il faut jouer et de combiner ça avec l'exploration arborescente, avant de faire jouer à la marchine des millions de parties contre elle-même. » Stanislas Dehaene, Yann Le Cun, Jacques Girardon, *La plus belle histoire de l'intelligence, Des origines aux neurones artificiels : vers une nouvelle étape de l'évolution,* Paris, Robert Laffont, 2018, p. 177.

9 *Ibid.,* p. 247.

10 Michel Bourdeau, *Pensée symbolique et intuition,* Paris, PUF, 1999, p. 12.

inconsciente, les connexionnistes semblent même avoir pour ambition de montrer que la majorité, voire l'ensemble, de nos facultés reposent sur l'intuition. Puisque leurs machines fonctionnent – et que leurs inventeurs s'accordent sur le fait qu'elles font preuve d'intuition – suffirait-il d'en décrire leur mécanisme pour que soient révélés les secrets de l'intuition ?

Avec le renouveau connexionniste, le projet d'intelligence artificielle s'invite donc au coeur de la philosophie, à ce que Derrida nomme « la gigantomachie philosophique autour de l'intuition et de l'intuitionnisme[11] ». Derrida, parcourant à grands pas l'histoire de la philosophie, montre comment, de Platon à Deleuze, en passant par Aristote, Plotin, Berkeley, Biran, Husserl et Bergson, la philosophie s'est caractérisée par un « intuitionnisme indéracinable qui constitue […] l'idée régulatrice de la philosophie même[12] ». Il y a « un intuitionnisme constitutif de la philosophie même, du geste qui consiste à philosopher[13] » : la rercherche d'une vision intuitive comme « contact », « toucher », ou « coïncidence » avec la vérité. Bien que l'étymologie du mot « intuition » le relie à la vision, il s'agit plutôt de *toucher,* le toucher renvoyant à la présence et à l'immédiateté[14]. L'intuition, écrit Derrida, au sujet de Bergson, est « cette reine », « une reine appelée à régner, et on voit les prétendants, mais aussi les *rivales,* se presser autour du trône : tous et toutes des concepts – des usurpatrices, des régicides[15] ! »

En proposant une définition de l'intuition, l'intelligence artificielle viendrait grossir le rang des « rivales » qui « se pressent autour du trône ». Portée par des ingénieurs plutôt que des philosophes, cette rivale s'appuye sur des textes d'un genre nouveau : du code s'effectuant sur des machines plutôt que des idées produisant des effets sur des humains. Elle s'éloigne de « l'intuitionnisme indéracinable » de la philosophie. Pour autant, elle n'est pas étrangère à une certaine tradition philosophique caractérisée par l'attraction vers les systèmes[16], voire les machines[17], que l'on trouve tout au long de l'histoire de la philosophie, en particulier chez Lulle, Leibniz, les logiciens, et certains courants de la « philosophie de l'esprit » comme l'approche

11 Jacques Derrida, *Le toucher, Jean-Luc Nancy,* Paris, Éditions Galilée, 2000.
12 Jacques Derrida, *op. cit.,* p. 144.
13 *Ibid.,* p. 138.
14 *Ibid.*
15 « Quel magnifique programme philosophique, et quelle scène, que d'images bien choisies, n'est-ce pas, autour de cette reine, l'intuition, une *psyché* en somme, une reine *appelée* à régner, et on voit les prétendants, mais aussi les *rivales,* se presser autour du trône : tous et toutes des concepts – des usurpatrices, des régicides ! » *Ibid.,* p. 141.
16 Jacques Bouveresse, *Qu'est-ce qu'un sytème philosophique ? Cours 2007 et 2008,* Paris, Collège de France, 2012.
17 Paolo Rossi, *Les philosophes et les machines, 1400-1700,* Paris, Presses Universitaires de France, 1996.

computationnelle (Fodor, Putnam…)[18]. Pour cette tradition, la référence à l'intuition est synonyme de manque de rigueur. Il vaut mieux, écrit Leibniz, « les règles d'Aristote et des Géomètres, comme, par exemple, de ne rien admettre (mis à part les principes, c'est-à-dire les vérités premières ou bien les hypothèses) qui n'ait été prouvé par une démonstration valable[19] ». Ils voient dans le thème de l'intuition une convergence de mythes (mythe de l'accès immédiat à la vérité, mythe d'un statut à part de l'humain…), et préfèrent valoriser le travail collectif d'élaboration de formes vérifiables.

Mais ces accusations peuvent être retournées à l'envoyeur. D'une part, négliger l'intuition revient également, d'une autre manière, à manquer de rigueur : c'est porter un discours sans prendre en compte les principes, les hypothèses et les vérités premières qui le soutiennent – en faisant comme si elles étaient évidentes pour tous – sous prétexte qu'ils se prêtent mal à la formalisation et à la vérification. D'autre part, il y a également une dimension mythique chez les tenants du projet d'intelligence artificielle et dans la tradition formaliste dont ils héritent : rêves de machines infaillibles, fantasmes de systèmes exhaustifs et univoques, mythes de la certitude, du monde et de la science comme machines…

Les prétentions du projet d'intelligence artificielle peuvent sembler excessives, mais celles de l'intuitionnisme philosophique le sont tout autant. D'un côté comme de l'autre, il y a une quête – vaine et suspecte – de la pureté. Pour l'intuitionnisme philosophique, cette quête se traduit par la recherche d'un accès *direct* à la vérité, sans qu'aucun média, voire aucun organe, n'en trouble la perception. Selon Platon, la *psyché* ne touche à la vérité que lorqu'elle « n'est troublée ni par la vue ni par l'ouïe, par aucun plaisir ou déplaisir du corps. Autrement dit quand elle a donné congé au sensible […][20]. » Derrida montre que cette expérience *sans corps* est une impossibilité et une contradiction : « il n'y a jamais d'expérience *pure et* immédiate du continu. Ni du proche. Ni de la proximité absolue[21]. » Nous verrons, sans surprise, que l'intelligence artificielle ne parvient pas à conquérir le trône de l'intuition. Mais cela est-il dû à une incapacité de l'intelligence artificielle ou bien à l'impossibilité, étant donné la manière contradictoire dont la philosophie intuitionniste décrit ce trône, d'y faire siéger quoi que ce soit ?

La pureté que poursuit le projet d'intelligence artificielle est différente. La recherche d'une description exhaustive des facultés cognitives aspire, en héritière de la tradition

18 Michael Esfeld, *La philosophie de l'esprit : Une introduction aux débats contemporains*, Paris, Armand Colin, 2012, chapitre 6, section 2.

19 G. W. Leibniz, « Réflexions sur la partie générale des Principes de Descartes (sur les articles 43, 45, 46) », in Lucy Prenant (ed.) *Œuvres de G. W. Leibniz*, Paris, Aubier-Montaigne, 1972, p. 297.

20 *Ibid.*, p. 138.

21 *Ibid.*, p. 144.

formaliste, à débarrasser le discours sur la pensée du vague, de l'implicite, de l'équivoque ; et, conformément à son postulat mécaniste, à assigner chaque événement à un faisceau de causes qui rende compte, *sans reste*, de chacun de ses aspects. L'exhaustivité attendue de la description ne laisse pas de place à l'ignorance, tandis que le formalisme du code est censé réduire l'équivoque de la description elle-même. Le but est d'éradiquer le bruit, aussi bien dans le langage que dans l'enchaînement des événéments.

L'intuitionnisme philosophique peut également s'interpréter comme une tentative, plus radicale encore, d'éradiquer le bruit : le meilleur moyen de se débarrasser de tout ce qui trouble nos supports d'accès à la vérité (media, organes) est peut-être de se débarrasser des supports eux-mêmes. Dans un cas, comme dans l'autre, il est postulé que le bruit dont il s'agit de se débarrasser *est second*. Il arrive après la vérité et les médias qu'il perturbe. L'opération de « nettoyage » présuppose que le bruit peut « s'enlever » : pour l'intuitionnisme philosophique, il existe une vérité non « troublée » ; pour le projet d'IA, il existe un modèle de l'esprit à découvrir et à décrire avec précision. En d'autres termes, la vérité ou le modèle seraient « sous » le voile du bruit et il suffirait d'une opération de nettoyage pour les révéler. Mais le bruit pourrait tout aussi bien être *premier* et la vérité – ou plutôt le sens, le terme restituant mieux son inscription dans le temps que la notion de vérité, trop souvent écartée du cours de l'histoire – seconde. C'est la thèse que nous allons défendre : le bruit, ou plus précisément le hasard, précèdent la constitution des ordres transitoires du monde, dont l'intelligence et le sens font partie.

Ce renversement (le hasard et le bruit précèdent l'ordre) a pour conséquence que toute tentative de description, ou de formalisation, est toujours *en deça* du monde. Nous pourrions dire que sommes *dépassés* par les événements, mais cela impliquerait que les événements arrivent *après nous*, or c'est l'inverse : nous n'arrivons qu'après les événements, avec un temps de retard, et cela vaut pour l'événement de notre intuition. Nous n'accédons à la cognition que longtemps après notre naissance. Surtout, nous restons débordés par le surgissement de nos idées. Après coup, chaque phénomène cognitif peut être décrit, mais nos descriptions ne peuvent anticiper sur les événements, elles ne peuvent aller plus vite que le temps. C'est donc une *hubris* que de vouloir, en se convaincant pour l'occasion de l'existence de « lois du monde » ou de la pensée, percer celles-ci à jour. Ce n'est pas une affaire, comme le veut un certain discours ambiant, de « complexité », mais bien de nouveauté. Pour le comprendre, et telle sera

notre tâche, il faut s'attarder sur ce qu'est le temps comme opérateur de métamorphoses, autrement dit sur ce que certains philosophes ont appelé « le devenir[22] ».

Les tenants de l'intuitionnisme philosophique, tout comme les chercheurs[23] en intelligence artificielle, s'égarent lorsqu'ils prennent leurs idées régulatrices (mythe de l'intuition immédiate et mythe du mécanisme universel) pour des réalités. Entre les deux mythes, il y a une voie étroite que nous tâcherons d'emprunter, en rendant justice à ce qui, dans chacune des démarches, est pertinent : la possibilité, en droit, de décrire tous les aspects de l'intelligence, et l'impossibilité, de fait, d'épuiser la question de la genèse des idées. Les deux démarches portent l'ambition de *décrire au mieux* ce qui advient lorsque l'on pense. À raison, le projet d'intelligence artificielle vient provoquer la philosophie : pourquoi, demandent les chercheurs, est-ce que certains aspects de l'intelligence ne seraient pas descriptibles ? Les égards de la philosophie vis-à-vis de l'intuition ne seraient-ils pas seulement le moyen de sacraliser la pensée et avec elle le travail du philosophe ? Peut-on encore sérieusement considérer que l'humain est, du fait de son intelligence, un membre « à part » du reste de la nature ? Mais, malgré la pertinence de la provocation, le projet d'intelligence artificielle ne semble pas en mesure de damer le pion à la philosophie intuitionniste et de mieux rendre compte de ce qui a lieu lorsque nous avons une idée. Décrire le projet d'intelligence artificielle depuis le point de vue de l'intuitionnisme philosophique, et inversement, nous permet ainsi de mieux situer et comprendre chacune des deux démarches en pointant leurs faiblesses respectives.

Nous allons décrire le projet d'intelligence artificielle dans sa version la plus récente, celle dite de l'« apprentissage profond » ou *deep learning,* et mettre cette description en regard avec une description de la notion d'intuition. Cela nous conduira à proposer une définition de l'intelligence artificielle, ainsi qu'une définition de l'intuition, afin d'avancer notre thèse : que les idées, comme l'ensemble des choses qui existent, surgissent ou naissent « sans raison », d'une contingence qui est première par rapport à l'instauration progressive de régularités qui passent ensuite, à tort, pour des « lois de la nature ». La question qui anime notre thèse est celle de la création et de la nouveauté. Si notre objet, le projet d'intelligence artificielle, relève du champ disciplinaire de l'histoire et de la philosophie des sciences et techniques, notre propos sera donc un propos de philosophie générale.

22 Gilles Deleuze , *Logique du sens,* Paris, Éditions de Minuit, 1969.

23 Nous utiliserons le masculin pour désigner les chercheurs en intelligence artificielle, ceux-ci étant et ayant été dans leur immense majorité des hommes. Selon les termes de Margaret Mitchell, les conférences rassemblent des « océans de mecs ». « Artificial intelligence has a sea of dudes problem », *Bloomberg,* 23 juin 2016, https://www.bloomberg.com/news/articles/2016-06-23/artificial-intelligence-has-a-sea-of-dudes-problem, page consultée le 10 décembre 2021.

Lorsque nous décrivons le monde, nos descriptions sont fragmentaires, tributaires de la singularité de notre point de vue et incapables de saisir la subtilité des métamorphoses de ce que nous percevons. Elles sont vite caduques. Pour autant, l'impuissance de nos représentations ne disqualifie pas l'effort de description. Au contraire, la tâche est toujours à recommencer, et nous nous devons de la recommencer pour moins mal saisir ce qui nous arrive, le jeu singulier qui caractérise notre époque, et renseigner les quelques générations qui, avant que notre espèce ne s'éteigne, voudront à leur tour comprendre quel est leur sort. La forme de l'humain n'a pas toujours existé, et elle disparaîtra. La philosophie s'est tournée avec obstination vers la lumière et l'éternel (établir des vérités, « éclairer » les causes et les idées…), mais c'est se mentir par omission quant à la nuit qui la précède, la dépasse, et qui finira par l'engloutir – sans cet abyme, nous n'aurions pas à penser.

Plan

Dans la première partie, nous allons étudier la pertinence des programmes d'intelligence artificielle comme mode de description de la pensée, et plus particulièrement de l'intuition. Pour ce faire, nous commencerons par présenter le projet d'intelligence artificielle en combinant plusieurs approches : le commentaire de certains textes fondateurs, notamment l'appel à projet de Dartmouth (1955-56) et l'article de Turing de 1950 ; une histoire succincte du projet d'intelligence artificelle, centrée sur celle des réseaux de neurones ; la description du fonctionnement et des principes de certains algorithmes, ainsi que de leur usage industriel. Cela nous permettra d'aboutir à une définition de l'intelligence artificielle et d'examiner la pertinence des dernières inventions en tant que modèles de certaines fonctions de l'esprit : la perception, l'induction, l'imagination et enfin l'intuition. Chemin faisant, nous verrons émerger un certain nombre de questions : l'ambition du projet d'intelligence artificielle est-elle de réaliser des machines intelligentes ou seulement des machines qui peuvent passer pour intelligentes – ou bien les deux reviennent au même ? Qu'est ce qu'un résultat valide en intelligence artificielle ? Selon quels critères peut-on dire d'une faculté de l'esprit qu'elle a été correctement simulée ? Existe-t-il une ou des qualités de l'esprit qu'aucun algorithme ne pourrait simuler ? La notion d'algorithme est-elle contradictoire avec celle de pensée ? Les machines peuvent-elles faire preuve de la même spontanéité que les humains ? Une entité non vivante peut-elle accéder à l'intelligence ? Les machines peuvent-elles être à l'origine de leurs

idées, comme semblent l'être les humains, ou bien l'originalité n'est-elle qu'une illusion ? Ces différentes questions invitent à se poser celle de l'intuition : existe-t-il une part originaire et spontanée de la pensée, qui serait, selon les opinions, impossible à représenter ou simuler – et donc une objection à l'ambition du projet d'intelligence artificielle – ou au contraire, son horizon, ce vers quoi le projet s'efforce d'aboutir ?

Dans la deuxième partie, nous proposerons une définition de l'intuition, en prenant pour point de départ la confusion qui règne à son sujet. Nous la décrirons « en creux », en nous appuyant sur une situation où elle est absente, le cas de la pensée dite « formelle » ou « aveugle ». Cela nous permettra de comprendre son rôle et d'aboutir à une définition de l'intuition qui mettra en évidence les carences de celle que proposent les chercheurs en intelligence artificielle. Par la même occasion, nous verrons comment le projet d'intelligence artificielle hérite d'une longue tradition visant à réduire l'intuition au profit de la pensée formelle, cette dernière étant conçue comme la seule à pouvoir garantir la rigueur et la certitude. Nous nous demanderons, à leur suite, s'il est possible de modéliser l'intuition ou, plus radicalement, de s'en passer complètement. En évitant à la fois une célébration excessive de l'intuition (pour son accès immédiat, sa pureté), ou au contraire une condamnation trop facile (pour son inconstance, sa variabilité, son opacité), nous nous attacherons à restituer au mieux les rôles respectifs de l'intuition et de la pensée formelle, puis nous nous demanderons si l'une peut exister sans l'autre, en particulier lorsqu'il s'agit de produire de nouvelles idées.

La troisième partie sera consacrée à la question de la création. De l'avis de Lee Sedol, AlphaGo a réussi plusieurs coups particulièrement inventifs. Pour ses concepteurs, le programme est plus créatif qu'un humain car, dépourvu de préjugés, il est en mesure d'explorer une plus grande palette de combinaisons. Loin d'être nécessaire à l'invention, notre subjectivité lui ferait obstacle. Nous examinerons cette idée en comparant AlphaGo avec des cas d'inventions et de découvertes ayant eu lieu à la faveur d'une suspension de l'effort conscient. Puis nous passerons au cas inverse, celui d'activités qui requièrent une reflexivité, en nous appuyant sur les analyses du jeu proposées par Roger Caillois, puis en commentant des textes de Bateson sur la conversation, et de Bachelard sur un certain type d'invention scientifique. Il s'agira d'interroger la capacité des humains et des programmes à traverser les crises, c'est-à-dire à inventer de nouvelles règles d'action pour faire face à une situation exceptionnelle. Selon les auteurs de l'appel à projet de Dartmouth, les machines pourront être rendues créatives, et donc capables de traverser des crises, pour peu qu'elles soient dotées du bon dosage de hasard. Nous verrons qu'une telle conception ne résiste pas à l'examen approfondi de la notion de hasard. En suivant les analyses de Clément Rosset, nous serons conduits à inverser la

perspective ordinaire : s'il y a du hasard, celui-ci doit être premier. Cela nous amènera ensuite, en nous aidant d'un texte de Bateson, à critiquer les notions de contrôle et de maîtrise, dont la popularité tient pour beaucoup à une forme de narcissisme. Le projet d'intelligence artificielle, en se donnant pour tâche de définir l'intelligence, est une manière de chercher à délimiter la forme de l'humain – par sa réussite, mais aussi par la vertu d'un éventuel échec qui donnerait à voir ce qui échappe à la machine. Qu'est-ce qui nous conduit à considérer qu'une entité est, sinon humaine, au moins doté de subjectivité ? Nous aborderons cette question d'une manière décalée, sous la forme d'une fiction mettant en scène un programmeur chargé de mettre au point un *chatbot* simulant la pensée de Jacques Derrida. Puis nous reprendrons le cours d'une réflexion philosophique plus classique en nous demandant s'il est possible, comme le pensent les chercheurs en intelligence artificielle, d'élucider la genèse de la pensée, ou s'il s'agit, comme le défendait Emil du Bois Reymond, d'une question qui dépasse les limites de la science. La *Critique de la raison pure* de Kant nous apportera une réponse et nous aidera à mieux situer les prétentions scientifiques du projet d'intelligence artificielle. Nous nous permettrons ensuite un moment de spéculation en considérant l'idée que l'intuition serait sans raison, ce qui nous amènera à revoir le rôle du projet d'intelligence artificielle – arc-bouté sur la recherche de modèles hors du temps et universalisables alors que sa pratique tend de plus en plus vers des dispositifs singuliers et dynamiques –, comme un mythe, et seul à même, en tant que mythe, de tenir un discours sur la genèse et le devenir. Nous verrons, pour conclure, en quoi ce mythe favorise une *obstination moderne,* c'est-à-dire, en faisant semblant de le nier – en échouant sans cesse à le nier –, ce mythe est l'occasion de reconduire l'arsenal théorique qui a fondé la colossale entreprise d'accaparement des ressources, des peuples, des autres espèces et des femmes par ceux qui se sont crus « modernes[24] » : les distinctions entre humain et non-humain, intelligent et non-intelligent, animé et inanimé.

24 Bruno Latour, *Nous n'avons jamais été modernes*, Paris, Éditions de la Découverte, 1991.

PREMIERE PARTIE

QU’EST-CE QUE L’INTELLIGENCE ARTIFICIELLE ?

1.1 Naissance du projet d'intelligence artificielle

1.1.1. La cybernétique

Bien que les fondateurs du projet d'intelligence artificielle aient cherché à s'en démarquer, on peut, et c'est l'idée que défend Jean-Pierre Dupuy, considérer que l'intelligence artificielle et les sciences cognitives descendent de la cybernétique[25]. Ce moment singulier de l'histoire des sciences prend sa source dans un article intitulé « Behavior, purpose and teleology », publié par Wiener, Bigelow et Rosenblueth en 1943[26]. Celui-ci propose une analyse comportementale qui s'applique à la fois aux machines et aux organismes vivants, avec des notions comme la « téléologie » (définie comme une finalité réglée par rétroaction négative) permettant d'étudier aussi bien les unes que les autres. Les perspectives ouvertes seront approfondies jusqu'en 1953, à la faveur d'un cycle de conférences interdisciplinaires financées par la fondation Macy. Si celles-ci sont animées et organisées par Warren McCulloch, Norbert Wiener en émerge comme le chef de file. Ses ouvrages touchent une large audience et popularisent sa vision de la cybernétique comme « science des *analogies* maîtrisées entre organismes et machines[27] ». Que l'on s'intéresse à un objet vivant ou inanimé, il peut être étudié à travers le prisme de notions communes comme l'organisation ou la communication. Si la cybernétique se veut un dialogue avec les sciences de la vie en général, les sciences de l'esprit demeurent en réalité son interlocuteur principal. En plus des mathématiciens, les participants aux conférences Macy sont psychologues, psychiatres, anthropologues, neurophysiologistes, linguistes[28]… Cet intérêt pour les sciences de l'esprit n'est pas le fruit du hasard. Prenant son essor pendant la guerre et l'immédiat après-guerre, le moment cybernétique est habité par l'idée que l'étude de l'humain, plus précisément l'application de la rigueur de la physique à la santé mentale des individus et des groupes, pourrait être le meilleur moyen de garantir enfin la paix[29]. Pour autant, il y a peu

25 Jean-Pierre Dupuy, *Aux origines des sciences cognitives*, Paris, La Découverte, 2005, p. 34-35.
26 Norbert Wiener, Arturo Rosenblueth, Julian Bigelow, « Behavior, Purpose and Teleology », *Philosophy of Science*, vol. 10, n° 1, 1943, p. 18-24.
27 Jean-Pierre Dupuy, *op. cit.*, p. 34-35.
28 Jean-Pierre Dupuy, *op. cit.*, p. 75-76. Un des rares représentants de la biologie, Max Delbruck, physicien étudiant les bactériophages, ne viendra qu'une seule fois pour repartir furieux.
29 *Ibid.*, p. 81.

de débat idéologique lors des conférences Macy. Les discussions tournent autour de certains problèmes pointus dont les solutions pourraient être utiles dans plusieurs disciplines. La cybernétique se présente avant tout comme une

> avant-garde de la démarche scientifique, tant par son objet – l'esprit, ce chef-d'œuvre de la création – que par ses concepts, ignorés jusque-là de la physique – la téléologie, l'information, la causalité circulaire, le feedback, etc. - et son style – réflexif, c'est-à-dire réfléchissant à l'emploi de ses outils conceptuels[30].

La cybernétique en tant que telle échoue à se constituer comme une discipline pérenne. Mais son vocabulaire et ses concepts (système, causalité circulaire, *feedback*, contrôle par la communication…) ont une postérité considérable dans de nombreux champs : psychologie, gestion, génétique, psychanalyse, anthropologie…

Malgré des divergences théoriques (voir section 1.3.7. L'IA contre la cybernétique), la cybernétique prépare le terrain pour le projet d'intelligence artificielle en familiarisant les scientifiques avec l'idée que l'esprit peut être conçu comme un processus sans sujet. Wiener soutient que « la volonté est de l'ordre du mécanisme » tandis que McCulloch fait de même « avec la perception, la pensée et la conscience[31] ». Conformément à l'analogie maîtresse du mouvement cybernétique, l'esprit peut s'étudier comme une machine – le « comme » étant entendu assez littéralement par certains membres. Ainsi McCulloch, dont le but affiché est de comprendre le fonctionnement du cerveau de manière à « savoir comment nous savons », affirme qu'il est « sur le point de concevoir le sujet connaissant [*knower*] comme une machine à calculer [*computing machine*][32] ». Puisqu'il y a analogie entre organismes et machines, l'engineering et la neurophysiologie pourraient ne faire qu'un. L'entreprise de McCulloch telle qu'il la conçoit est une enquête sur les conditions de possibilité formelles et matérielles de toute connaissance, humaine ou non. De là à envisager la fabrication de machines simulant l'esprit, il n'y a qu'un pas, que McCulloch semble prêt à faire, puisqu'il assiste au séminaire de Dartmouth, moment fondateur du projet d'intelligence artificielle. Surtout, il lui fournit un de ses premiers outils en théorisant, avec Walter Pitts, les réseaux de neurones (voir infra 1.2.2.). En contribuant à diffuser l'idée d'une analogie entre cerveau (ou esprit) et machine, le mouvement cybernétique aide à diffuser celle de la construction de machines intelligentes. Autour de 1950, lorsque les médias s'intéressent aux conférences Macy, ils présentent le groupe

30 *Ibid.,* p. 157.
31 *Ibid.,* p. 118.
32 *Ibid.,* p. 53.

cybernétique comme « ayant démontré que le cerveau est une machine », ce qui permet, selon les journalistes, d'« envisager de construire des machines intelligentes[33] ».

1.1.2. Invention de l'ordinateur : un cerveau électronique ?

L'ordinateur naît à la faveur de l'effort scientifique et industriel sans précédent qu'occasionne la seconde guerre mondiale. Les pays belligérants entretiennent des projets de calculateurs aux fonctions diverses : cryptographie, balistique… Dans un document préparatoire à l'un de ces projets[34], John von Neumann opère la synthèse des différentes recherches et fait un certain nombre de choix. Il prend position en faveur d'opérations binaires (plutôt que décimales), de composants entièrement électroniques (tubes à vides plutôt que relais électromécaniques) et d'une architecture stockant les programmes dans la même mémoire que les données. Les choix opérés par von Neumann font référence et sont adoptés comme le standard de l'ordinateur moderne – ainsi que le remarque Alan Turing dans son propre rapport pour un ordinateur pilote en Angleterre[35]. Dans son rapport, von Neumann décrit l'ordinateur *au moyen d'une analogie avec l'esprit et le cerveau humain.* C'est une métaphore scientifique[36] au sens où le cerveau est *l'objet connu* qui permet d'appréhender l'*objet inconnu* qu'est l'ordinateur naissant[37]. On dira de lui qu'il effectue des « opérations » et exécute des « instructions » de la « mémoire ». Surtout, von Neumann compare les « éléments » de base du calculateur avec des « neurones ».

Von Neumann ne cache pas qu'il ne s'agit que d'une analogie[38] et qu'il s'agit – comme toute analogie – de ne retenir que *quelques éléments communs* entre le terme comparant et le terme comparé. En prenant exemple sur McCulloch et Pitts, il déclare ne retenir des neurones que leur aspect « tout ou rien », c'est à dire leur similarité avec des opérateurs logiques

33 *Ibid.*, p. 89. Les participants aux conférences Macy réagissent avec mépris aux déclarations des journalistes.

34 John von Neumann, « First Draft of a Report on the EDVAC » (1945), *IEEE Annals of the History of Computing*, vol. 15, No. 4, 1993, p. 27-43.

35 Alan Turing, « Proposals for Development in the Mathematics Division of an Automatic Computing Engine (ACE) » (1945), Com Sci 57, National Physical laboratory, Teddington, UK, 1972.

36 Nous reprenons la formulation de David West et Larry Travis, qui utilisent la définition de la métaphore par MacCormac comme utilisation des propriétés d'un objet connu pour décrire un objet inconnu auquel il ressemble : E. R. MacCormac, « Scientific Metaphors as Necessary Conceptual Limitations of Science. » *in* N. Rescher (dir.) *The Limits of Lawfulness*, Pittsburgh, University of Pittsburgh Center for the Philosophy of Science. 1983, p. 185–203. Cité par David West et Larry Travis, « The Computational Metaphor and Artificial Intelligence : A Reflective Examination of a Theoretical Falsework », *AI Magazine*, 12, Janvier 1991, p. 64-79.

37 « The computer was the strange object in the metaphoric relationship, and the known object was mind. » David West and Larry Travis, *op. cit.*

38 Le titre de la partie 4.0 est « neuron analogy ».

booléens. Tous les autres aspects « plus compliqués » du fonctionnement des neurones, sont mis de côté[39].

Dans un ouvrage ultérieur[40], von Neumann revient plus en détail sur les tenants et les aboutissants de la comparaison entre cerveau et ordinateur. Il souligne ses limites en remarquant que les fonctions cognitives de haut niveau semblent dépendre de l'organisation matérielle du cerveau (remarque qui serait aujourd'hui qualifiée de connexionniste). Aussi faudrait-il probablement, pour qu'une machine pense, qu'elle ait la même l'architecture que le cerveau. Mais ces précautions ne font guère recette, l'analogie résonne trop avec l'air du temps. Pour les scientifiques (à commencer par Turing[41]), la presse et le public, les ordinateurs naissants sont bien des « cerveaux électroniques ». Depuis, l'analogie avec le cerveau s'est si bien intégrée au langage courant qu'il est difficile d'imaginer par quel autre moyen décrire un ordinateur. Pourtant, l'histoire aurait pu prendre un autre cours. Un siècle plus tôt, quand Babbage conçoit les prémices de l'ordinateur, il emploie le champ lexical des usines de son temps : le stockage est un « entrepôt » (« store ») et le calcul est effectué par une « usine » (« mill »)[42]. Le plan de sa « machine à différences » s'inspire des usines (en l'occurence de l'« usine de calcul » de Gaspard de Prony[43]) et non d'un hypothétique fonctionnement de la pensée.

1.1.3. Retournement : l'ordinateur permet-il de comprendre le cerveau ?

Il suffira de quelques années pour que l'ordinateur devienne un objet familier et que *la comparaison avec l'esprit s'inverse*. Les ordinateurs sont abondamment évoqués par la presse et les ouvrages de vulgarisation. On en trouve dans les laboratoires, puis dans les entreprises et les écoles. Il est mis en scène dans les romans et les films de science-fiction. Le langage courant ayant gardé les traces de l'analogie employée pour le décrire au moment de son invention, celle-

39 « Following W.S. MacCulloch and W. Pitts ("A logical calculus of the ideas immanent in nervous activity," Bull. Math. Biophysics, vol. 5 (1943), p. 115–133) we ignore the more complicated aspects of neuron functioning: Thresholds, temporal summation, relative inhibition, changes of the threshold by after-effects of stimulation beyond the synaptic delay, etc. » John von Neumann, *op. cit.*

40 John von Neumann, *L'ordinateur et le cerveau,* Paris, Champs Flammarion, 1996.

41 Andrew Hodges, *Alan Turing ou l'énigme de l'intelligence*, Paris, Payot, 2004, (traduit de l'anglais : *Alan Turing, The Enigma*, Princeton, Princeton University Press, 2014), p. 148.

42 « It should be noted, however, that Babbage used metaphors of mills and stores derived from the mechanistic technology of his era rather than mental metaphors. » David West and Larry Travis, *op. cit.*

43 Voir Margaret Bradley, « Gaspard-Clair-François-Marie Riche de Prony (1755-1839), Constructeur de ponts », in *Bulletin de la SABIX,* « Regards sur des carrières de polytechniciens au XIXe siècle », 48 | 2011.

ci est alors mise à profit dans l'autre direction : *c'est l'ordinateur, objet bien connu, qui devient le moyen de décrire l'esprit, objet mal connu*[44]. En familiarisant le public avec le rapprochement entre les deux entités, la première métaphore (parler de l'ordinateur comme d'un esprit) a préparé le terrain pour l'apparition de la deuxième (parler de l'esprit comme d'un ordinateur)[45]. Si bien qu'aujourd'hui *l'objet à connaître* en passant par la comparaison avec autre chose, est la pensée. La pensée, en tant qu'elle nous est la plus proche, la plus familière, aura servi à appréhender l'arrivée d'un objet étranger – l'ordinateur. Mais une fois ce dernier intégré, la pensée apparaît paradoxalement comme *ce qui est le moins connu.*

Cependant, dès la naissance de l'ordinateur, plusieurs représentants de la théorie de la *Gestalt* émigrés aux Etats-Unis (Lewin, Wertheimer, Köhler…) entrent en discussion avec le mouvement cybernétique (Köhler publie une critique de *Cybernetics* de Wiener[46], Lewin participe à plusieurs des conférences Macy…) et critiquent l'idée que les ordinateurs et les machines en général puissent servir de modèles pour le cerveau ou la pensée. Les machines, disent-ils, ne sauraient rendre compte de notre capacité à percevoir le sens d'un problème (*insight*)[47]. Ils montrent, par des expériences et des observations sur les singes, les poulets ou les embryons, que cette faculté d'*insight* n'est pas propre à l'esprit humain, c'est une propriété des organismes en général. Pour comprendre l'*insight,* il faut considérer l'organisme du point de vue de son ensemble et non, comme le fait la comparaison avec la machine, du point de vue de ses composants ou de ses fonctions, l'organisme comme totalité ayant la capacité, lorsqu'il est tendu vers un but, de requisitionner et détourner n'importe lequel de ses organes (ou de ses fonctions) afin de l'atteindre. Notre capacité à trouver la solution d'un problème ne se logeant dans aucune fonction en particulier, les machines échouent à en rendre compte. Tout à son optimisme, McCulloch fait d'abord peu de cas de cette critique. Mais, une dizaine d'années plus tard, il admet que « le problème de l'insight, ou intuition, ou invention – appelez-le comme vous voulez, nous ne le comprenons pas[48] ».

---

44 « The commonly adopted but scientifically erroneous notion that the mind was something we understood well enough to use as a metaphor for the unknown computer faded even though the mental metaphors applied to computers persisted. Instead, the computer became the metaphor of choice for a renewed research effort directed toward understanding the mind. », David West et Larry Travis, *op. cit.*, p. 71.

45 « In large part, this status came about because the distance between computers and minds as dissimilar entities had already been bridged and significantly reduced by the first set of metaphors that related minds to computers. Those hearing the new metaphors were predisposed to accept them. » *Ibid.*

46 Wolfgang Köhler, « Review of *Cybernetics or Control and Communication in the Animal and the Machine* by Norbert Wiener », *Social Research,* 18, 1951, p. 128, cité par David Bates, « Creating Insight : Gestalt Theory and the Early Computer » in Jessica Riskin (ed.), *Genesis Redux, Essays in the History and Philosophy of Artificial Life,* Chicago, The University of Chicago Press, 2007, p. 240.

47 David Bates, « Creating Insight : Gestalt Theory and the Early Computer », *op. cit.,* p. 238-239.

48 « the problem of insight, or intuition, or invention – call it what ou will – we do not understand. » Warren McCulloch, « What Is a Number, That a Man May Know It, and a Man, That He May Know a Number ? » in

1.1.4. Turing : un ordinateur peut-il penser ? (1950)

En 1950, Turing publie un article[49] qui préfigure ce revirement puisque la machine y est déjà l'objet bien défini permettant d'approcher l'objet inconnu qu'est la pensée. Après avoir formulé la question qui l'occupe (« Je propose de considérer la question : 'Les machines peuvent-elles penser ?'[50] »), Turing remarque aussitôt qu'« il faudrait commencer par définir le sens des termes 'machine' et 'pensée'[51] ». Ces définitions sont tributaires de l'opinion commune, or il faudrait éviter que la réflexion tourne au « sondage d'opinion ». Voilà pourquoi il vaut mieux « remplacer la question par une autre » plutôt que de s'attarder à définir les termes.

> Les définitions peuvent être conçues de manière à refléter, autant que possible, l'utilisation normale des mots, mais cette attitude est dangereuse. Si on doit trouver la signification des mots 'machines' et 'penser' en examinant comment ils sont communément utilisés, il est difficile d'échapper à la conclusion que la signification de la question « Les machines peuvent-elles penser ? » et la réponse à cette question doivent être recherchées dans une étude statistique telle que le sondage d'opinion. Mais cela est absurde. Au lieu de m'essayer à une telle définition, je remplacerai la question par une autre, qui lui est étroitement liée et qui est exprimée en des termes relativement non ambigus[52].

Pourquoi une telle remarque en introduction de son article ? S'il est compréhensible que Turing manifeste une volonté de s'affranchir des ambiguïtés de l'opinion commune, il est curieux qu'il ne mentionne pas la possibilité de recourir à la formalisation. Lui qui a apporté une définition formelle à la notion d'algorithme, n'évoque pas l'éventualité de répéter la prouesse accomplie en 1936[53]. Fait encore plus étrange, en ce qui concerne la machine, sa remarque est une prétérition. Quelques pages plus loin, Turing annonce en effet : « La question que nous avons

---

Embodiments of Mind, Cambridge MA, MIT Press, 1965, p. 14, cité par David Bates, « Creating Insight : Gestalt Theory and the Early Computer », *op. cit.*

49 Alan Turing, « Computing Machinery and Intelligence », *Mind*, vol. 59, No. 236, Oct. 1950, p. 433-460. Nous utilisons la version française : Alan Turing, « Les ordinateurs et l'intelligence » in Alan Turing et Jean-Yves Girard, *La machine de Turing*, Paris, Éditions du Seuil, 1995.

50 Alan Turing, *op. cit.*, p. 135.

51 *Ibid.*

52 *Ibid.*

53 Alan Turing, « On Computable Numbers, with an Application to the Entscheidungsproblem », in Jack 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, 2004, p. 58-97.

posée dans la section 1 ne sera totalement définie que lorsque que nous aurons spécifié ce que le mot 'machine' signifie. » Il détaille quelles machines peuvent « prendre part à notre jeu », définit ce qu'est un ordinateur et ce qu'est une machine universelle[54]. Le terme de « machine » étant ainsi défini sans avoir eu recours au « sondage d'opinion », la remarque inaugurale de Turing ne s'applique donc qu'au seul terme de « penser », terme ambigu et tributaire de l'opinion commune, qui justifie le « remplacement » de la question par un jeu bien défini.

L'inconfort que manifeste Turing renvoie à un *topos* philosophique : la difficulté à séparer *ce qu'est la pensée* de *ce qu'en dit l'opinion commune,* autrement dit à s'affranchir de la *doxa*. Si Turing a pris le temps de définir clairement ce qu'est une machine tout en maintenant qu'il vaut mieux « remplac[er] la question par une autre » – qui ne fasse appel à aucune définition de la pensée – c'est donc qu'il considère qu'il est vain de chercher à *formaliser ce qu'est la pensée* pour la distinguer de l'opinion commune. Le « jeu de la pensée » a des règles trop floues, il vaut mieux le remplacer par un autre.

À l'ambiguïté des définitions communes s'ajoute le fait qu'elles évoluent en fonction de l'époque. Pour Turing, l'évolution de l'usage des mots est à la fois ce qui empêche d'apporter une réponse à la question et *ce qui permettra finalement d'y répondre, par l'affirmative* : « je crois qu'à la fin du siècle l'usage, les mots et l'éducation de l'opinion générale auront tant changé que l'on pourra parler de machines pensantes sans s'attendre à être contredit[55].» L'opinion commune évolue, et avec elle la signification des mots. Il y a une *contingence du sens des mots* et plus spécifiquement une contingence du sens donné au mot « penser ». Pour Turing, c'est la même contingence du sens qui *empêche de poser la question aujourd'hui* et qui *l'amènera demain à être considérée comme résolue*.

Bien que Turing soit prêt à parier que la question sera résolue par l'évolution de la *doxa,* il persiste à vouloir la remplacer « par une autre » qui ne repose pas sur ces termes aux définitions ambiguës et changeantes. Pour ce faire, il choisit de remplacer la question par un test. Les « termes relativement non ambigus » de la nouvelle question consistent dans le fait qu'elle est une épreuve pratique :

> Au lieu de m'essayer à une telle définition, je remplacerai la question par une autre, qui lui est étroitement liée et qui est exprimée en des termes relativement non ambigus. Le

54 Alan Turing, « Les ordinateurs et l'intelligence », *op. cit.,* p. 138-148.
55 *Ibid.,* p. 149.

> problème reformulé peut être décrit dans les termes d'un jeu que nous appellerons le 'jeu de l'imitation'[56].

Puisque le sens des mots est contestable, mieux vaut remplacer la question par une épreuve pratique, autrement dit se prémunir de la contingence du sens en *se passant de ces mots* trop changeants pour *se fier aux choses.* Turing s'inspire du « jeu de l'imitation » consistant à placer un homme et une femme dans des pièces séparées. Depuis une troisième pièce, un jury leur pose des questions par écrit et tente de distinguer l'homme de la femme. Dans le « test » que propose Turing, c'est un ordinateur qui remplace l'homme. Le jury est mis au défi de reconnaître lequel de ses interlocuteurs est une machine. Après avoir décrit le jeu, Turing réaffirme que celui-ci *remplace* la question initiale :

> Nous posons maintenant la question : 'Qu'arrive-t-il si une machine prend la place de A dans le jeu ? L'interrogateur se trompera-t-il aussi souvent que lorsque le jeu se déroule entre un homme et une femme ?' Ces questions remplacent la question originale : 'Les machines peuvent-elles penser[57] ?'.

Admettre avec Turing que le jeu de l'imitation est un remplacement pertinent de la question initiale (« Les machines peuvent-elles penser ? ») revient à admettre que le résultat du jeu apporte une *réponse* indirecte à la question. En proposant une réponse qui nous aide à trancher quant à la possibilité, pour les machines, de penser, le jeu de l'imitation prétend être un instrument de découverte pour investiguer ce que sont les machines et ce qu'est la pensée. Dans la mesure où, des deux termes que le dispositif vise à rapprocher, la machine est clairement définie, et la pensée ne l'est pas, on en déduit que l'enjeu du test est *de mieux définir la pensée en s'aidant pour cela de notre connaissance de la machine.* Autrement dit, il s'agit de prendre le fonctionnement des machines comme un moyen *d'approximer* ce qu'est la pensée.

Mais le statut de cette approximation n'est pas clair. Si le jeu de l'imitation permet *indirectement* de répondre à la question inaugurale (« Les machines peuvent-elles penser ? »), est-ce à dire qu'il apporte une connaissance quant à la pensée ? S'agit-il de philosophie par la bande ? Et de quelle nature est cette connaissance que se passe de mots, qui se targue d'être *indépendante du sens que leur donne l'époque ?*

Il y aurait là une ambition considérable du jeu de l'imitation, et dans sa suite du projet d'intelligence artificielle : faire de la philosophie en s'affranchissant des contingences du sens.

56 *Ibid.,* p. 135.
57 *Ibid.,* p. 135-136.

Serait-il possible de *réaliser la philosophie* : en faire une *chose* en remplaçant les questions par des dispositifs techniques, et ainsi l'émanciper de l'esprit du temps et des contingences du langage ?

1.1.5. Le *credo* de Turing (1950)

Avant de présenter ses arguments, Alan Turing annonce qu'il va présenter son opinion (« Cela simplifiera les choses pour le lecteur si j'expose d'abord mes propres vues sur le sujet ») et se lance dans une profession de foi :

> **Je crois** que la question originale 'Les machines peuvent-elles penser ? » a trop peu de sens pour mériter une discussion. Néanmoins, **je crois** qu'à la fin du siècle l'usage, les mots et l'éducation de l'opinion générale auront tant changé que l'on pourra parler de machines pensantes sans s'attendre à être contredit[58].

La répétition du « je crois » est remarquable – et remarquée par Turing qui justifie aussitôt l'exposé de ses croyances par un nouveau « je crois » :

> **Je crois** de plus qu'il ne sert à rien de dissimuler **ces croyances**. L'idée populaire selon laquelle les savants avancent inexorablement d'un fait bien établi à un autre, sans être influencés par des hypothèses non vérifiées, est absolument fausse. Pourvu que nous sachions clairement quels sont les faits prouvés et qu'elles sont les hypothèses, aucun mal ne peut en résulter. Les hypothèses sont de grande importance puisqu'elles suggèrent d'utiles voies de recherches[59].

Turing ne se contente pas de formuler ses conjectures (comme le feront ensuite les organisateurs de la conférence de Dartmouth) ou d'asséner un argument d'évidence (« Nous tenons ces vérités pour évidentes par elles-même »). Il préfère admettre avec franchise l'absence de fondement des hypothèses. En cela il manifeste non seulement qu'il croit en la réalisation de machines pensantes, mais aussi que pour croire à la réalisation de machines pensantes, il faut croire à ses intuitions, et croire à l'intuition en général, c'est-à-dire à l'utilité de propositions non vérifiées.

58 *Ibid.,* p. 149. Nous soulignons.
59 *Ibid.*

Turing croit en l'intelligence artificielle et assume le fait que *croire à l'intelligence artificielle implique de croire en l'intuition.*

Ensuite, Turing se comporte comme un Père de l'Église. Plutôt que de donner ses propres arguments, il préfère répondre à ceux qui ont des croyances opposées. Le voilà qui réfute une à une les objections qui pourraient lui être faites, comme si l'évidence était telle que la charge de la preuve devait revenir à l'autre partie. Si Turing avance de sérieux arguments à la faveur de ces réfutations, il faut souligner la bizarrerie de la tournure qui revient à préférer l'absence de réfutation solide à une démonstration valide. Turing veut établir que l'intelligence d'une machine qui ressemble suffisamment à un être humain est *difficile à réfuter, voire irréfutable.* Mais qu'elle soit irréfutable ne signifie pas qu'elle soit *prouvée.* Ne pas pouvoir réfuter l'intelligence d'une machine capable de tromper un jury est-il suffisant pour prouver l'intelligence de cette machine ?

À nouveau, Turing le reconnaît avec franchise : « Le lecteur aura compris que je n'ai pas d'argument positif très convaincant pour soutenir mon point de vue. Si j'en avais, je n'aurai pas pris tant de peine à montrer les erreurs des points de vue opposés aux miens[60]. » Puis il annonce qu'il va tout de même avancer des arguments : « Les preuves que j'ai, je vais maintenant les donner[61]. » Mais en guise de preuves, Turing se contente de mentionner deux nouvelles analogies. La première compare l'esprit humain et une pile atomique. Certaines idées, lorsqu'elles rencontrent certains esprits, déclenchent une sorte de réaction en chaîne. « Une idée proposée à un tel esprit pourra donner lieu à l'apparition de toute une 'théorie' constituée d'idées secondaires, tertiaires, ou encore plus éloignées[62]. » Si nous pouvons mettre au point une pile atomique, pouvons-nous fabriquer une machine capable de tels cataclysmes philosophiques ? Turing ne fait que poser la question et passe aussitôt à la deuxième analogie. Les fonctionnalités de l'esprit sont comme les couches d'un oignon que l'on épluche : si l'on enlève chacune des fonctions de l'esprit s'expliquant en termes purement mécaniques, arrive-t-on à « l'esprit réel » ou « arrivons-nous finalement à la peau qui ne contient rien[63] » ? Avec ces deux paragraphes, Turing ne fournit pas vraiment de « preuves ». Il ne fait que reformuler la question initiale en s'appuyant sur des analogies : l'enchaînement des idées est comparé à une réaction nucléaire, et les fonctions de l'esprit sont comparées aux peaux d'un oignon. Là encore, il l'avoue honnêtement : « Ces deux derniers paragraphes ne prétendent pas être des arguments

60 Alan Turing, *op. cit.*, p. 166.
61 *Ibid.*
62 *Ibid.*, p. 167.
63 *Ibid.*

convaincants. On les décrirait mieux en disant que ce sont des 'déclamations tendant à produire une croyance'[64]. » Turing a une intuition. Il *croit* qu'à la fin du siècle nous pourrons parler de machine pensante sans être contredit[65]. Mais il n'a aucun argument à l'appui de cette intuition. Il doit se contenter de réfuter ceux qui ne partagent pas sa croyance, et de quelques analogies. Après avoir présenté ces dernières, il les discrédite pour revenir au test : « Le seul élément vraiment satisfaisant qui puisse soutenir le point de vue exprimé au début de la section 6 nous sera fourni par la réalisation, à la fin du siècle, de l'expérience décrite[66]. » En d'autres termes, ce que Turing présente comme « le seul élément vraiment satisfaisant » est la fabrication d'une machine qui passe le test.

1.1.6. Anticiper ou participer

Revenons à l'analogie que Turing propose entre l'esprit humain et une pile atomique. Comme des boules de billard, les idées viendraient heurter nos esprits et y créer un effet limité : « On pourrait dire qu'un homme peut 'injecter' une idée dans la machine, laquelle réagira jusqu'à un certain point, puis retournera à l'immobilité, comme une corde de piano frappée par un marteau[67]. » C'est ainsi que nous fonctionnerions la plupart du temps : « Une idée proposée à un tel esprit donnera lieu, en moyenne, à l'apparition de moins d'une idée en réponse[68]. » Mais certaines idées, communiquées à certains esprits, y produisent une sorte de réaction en chaîne : une idée en entraîne une autre, et chaque nouvelle idée s'amplifie. « Une idée proposée à un tel esprit pourra donner lieu à l'apparition de toute une 'théorie' constituée d'idées secondaires, tertiaires, ou encore plus éloignées[69]. » Assurément, l'idée du test de Turing, en commençant par l'effet qu'elle fait sur l'esprit de Turing lui-même, fait partie de ces « bombes » intellectuelles. Elle y produit des idées « secondaires, tertiaires », mais surtout très éloignées : l'article saute des jeux de devinettes à la pile atomique, en passant par la poésie et la télépathie. La bizarrerie du texte de Turing est soulignée par Bruno Latour :

64 *Ibid.*
65 Il faut souligner que Turing ne précise pas si croire que « à la fin du siècle, nous pourrons parler de machine pensante sans être contredit » est équivalent à croire que « à la fin du siècle, il y aura des machines pensantes ».
66 *Ibid.,* p. 167-168.
67 *Ibid.*
68 Alan Turing, *op. cit.,* p. 167.
69 *Ibid.*

> Imaginez ce qu'on dirait d'un informaticien, d'un spécialiste de robotique, d'un psychologue ou d'un neurobiologiste qui enverrait à une revue fort sérieuse un long article comprenant les historiettes suivantes : un jeu de rôle dans lequel un homme, caché derrière un paravent, cherche à se faire prendre pour une femme ; la description kafkaïenne du travail d'un malheureux bureaucrate noircissant des rames de papier sans jamais lever le nez de son Code ; une histoire d'ingénieurs, tous du même sexe, essayant de cloner un humain à partir d'une « *seule cellule de sa peau* » ; une Maman qui demande à Tom « *de passer chaque matin chez le cordonnier* » ; un démon de Laplace enfermé dans une machine pour échapper aux effets des théories du chaos ; un bref roman d'anticipation sur les conséquences intellectuelles des innovations techniques ; une digression sur l'âme des femmes dans la théologie musulmane ; une autre sur la transmigration ; encore une autre sur le droit des humains de servir d'instruments à la volonté de Dieu en « *offrant une demeure pour les âmes qu'Il a créées* » ; un petit dialogue à propos du sens exact d'un sonnet ; un long passage sur la télépathie et autres phénomènes extrasensoriels ; un compte rendu de l'éducation très cruelle reçu par un enfant auquel on inculque ses leçons par de grands coups sur la tête… pour finir par la Constitution des États-Unis et la stupéfaction du programmeur qui voit la machine faire des choses qu'il n'avait pas prévues bien qu'il en ait écrit chaque ligne de code[70] ?

L'idée du test provoque chez Turing un excès d'imagination qui contredit ce qu'il cherche à montrer : les idées ne sont pas seulement « éloignées », elles sont désordonnées et parfois contradictoires. Il est difficile de voir en quoi la question des machines pensantes devrait *nécessairement* entraîner des réflexions sur le test, et celles-ci amener au clonage ou à la télépathie. Turing semble croire que les idées se succèdent comme l'effet succède à la cause, autrement dit que la pensée se déroule de manière nécessaire, mais son propos illustre le contraire. Les arguments ne s'enchaînent pas logiquement. Ils sautent d'un domaine à un autre, voire se contredisent. Cela est manifeste si l'on brosse à grands traits les principales étapes de l'article : Turing annonce qu'il traitera de la question « les machines peuvent-elles penser ? », puis, à cause de la contingence du sens de « machine » et « pensée », choisit de « remplacer la question par une autre », celle du test. Se contredisant ensuite, il définit clairement ce qu'il entend par « machine ». Puis, sans donner d'arguments en faveur de sa position, il répond aux objections qui pourraient être formulées à l'encontre du test, ou plus généralement de la possibilité de machines pensantes. Ce sont sept paragraphes très hétérogènes, dont certains passages rappellent *Alice au pays des merveilles* (peut-on comparer une personne à un jour

70 Bruno Latour, *Chroniques d'un amateur de sciences,* Paris, Presses des Mines, 2006, p. 63-65.

d'hiver ?), tandis que d'autres versent dans la parapsychologie (l'évidence statistique en faveur de la télépathie serait « accablante », écrit Turing). Après avoir commenté chacune des objections, Turing concède qu'il « n'a pas d'argument positif très convaincant pour soutenir [son] point de vue » et annonce qu'il va donner des preuves. Ce qu'il ne fait pas. Au lieu de cela, il présente deux analogies tout aussi incongrues : une pile atomique et un oignon sont censés convaincre le lecteur du bien-fondé de son *credo.* Enfin, il termine par plusieurs pages de considérations sur la manière d'enseigner aux machines à penser.

Alors que l'article argumente en faveur d'une certaine nécessité de la pensée, il le fait d'une façon qui manifeste, au contraire, toute la bizarrerie, la fécondité, et la contingence de l'imagination au travail.

> Rien de moins formaliste, rien de plus charnel, bizarre, hésitant, incohérent, multiple, que cette exploration à tâtons d'un monde qui n'existe pas encore et dont Turing dessine la baroque esthétique. Pour peu qu'on se donne la peine de le relire, le texte original déploie tout un opéra. Jamais l'invention littéraire, l'imagination la plus débridée, l'audace intellectuelle, les pièges tendus par l'inconscient, ne se sont mêlés si intimement à l'invention technologique, à la métaphysique et au formalisme logique[71].

En amont du texte, au paragraphe 6, Turing répond à « l'objection de Lady Lovelace » en remarquant que, si nous pouvons être surpris par ce que font les machines, comme nous sommes surpris par ce que disent ou font les humains, c'est parce que nous ne prenons pas le temps de calculer minutieusement les conséquences de ce qui se présente à nous. Les idées qui nous viennent, comme les étapes que suivent les machines, s'enchaînent selon une certaine logique. Il suffit, à condition d'en avoir la puissance de calcul, de suivre cette logique pour pouvoir les anticiper.

Bien malgré lui, Turing se fait l'exemple de tout autre chose : ce n'est pas une question de puissance de calcul mais de trajectoire. Si ses réflexions échouent à anticiper sur le devenir des machines, ce n'est pas parce qu'elles ne vont pas assez vite, c'est qu'il semble y avoir une infinité de chemins. Ce n'est pas tant que le « réel » aille trop vite, c'est qu'il foisonne dans trop de directions hétérogènes. La pensée échoue à l'anticiper. Ce foisonnement la dépasse, et elle-même y participe. La pensée prend trop de directions simultanées et ces directions participent au buissonnement des événements. Comment pourrait-elle anticiper quoi que ce soit si elle se révèle incapable d'anticiper sur ses propres chemins ?

71 Bruno Latour, *Ibid.*

Pour conclure l'article, Turing s'interroge : vaut-il mieux mettre une machine aux échecs ou bien l'équiper « avec les meilleurs organes sensoriels que l'on puisse acheter, puis lui apprendre à comprendre et à parler[72] » ? Turing esquisse là plusieurs méthodes, ou plusieurs chemins, qui seront, selon sa recommandation (« je pense qu'il faudrait essayer les deux voies »), tous deux empruntés par ses successeurs. Aussi ne conclut-il pas avec *une réponse*, mais avec une, ou plus exactement *plusieurs méthodes*. La question « les machines peuvent-elles penser ? » n'aura pas reçu de réponse. Elle aura été l'occasion d'un foisonnement, d'un désordre, qui, sans permettre de mettre ses pensées en ordre, les auront toutefois mises *en ordre de marche*, comme en témoignent les derniers mots de l'article, infiniment plus prudents, plus pragmatiques, que l'introduction : « Notre vision de l'avenir est limitée, mais du moins nous voyons qu'il nous reste bien des choses à faire[73]. » Avec cette conclusion, la pensée de Turing ne prétend plus *anticiper* sur un ordre des choses prédéterminé (les machines pourront-elles penser en l'an 2000 ?), mais *participer* à leur foisonnement dans des trajectoires multiples qui, sans forcément déroger à la causalité, déjouent en tout cas la prévisibilité et amènent le sujet (ici, Turing), à être surpris, légèrement débordé, aussi bien par les événements que par ce qu'il en pense.

1.1.7. La conjecture de Dartmouth (1955-1956)

En 1955, John McCarthy, Marvin Minksy, Claude Shannon et Nathaniel Rochester présentent leurs intentions de recherche dans un document de quelques pages[74] visant à faire financer un séminaire d'été à Dartmouth. C'est la première fois qu'est utilisé le terme « intelligence artificielle » et que des chercheurs aux intérêts divers se réunissent sous cette bannière. La conférence de Dartmouth peut être considérée comme le coup d'envoi de l'intelligence artificielle en tant que champ scientifique. C'est là que le projet est nommé, qu'est formulée le plus clairement son hypothèse directrice, et que se rassemblent une diversité de chercheurs dont certains essaiment dans des laboratoires en se réclamant de l'intelligence artificielle[75]. Dès les premières lignes du document, les auteurs mentionnent l'hypothèse directrice du séminaire :

72 Alan Turing, *op. cit.*, p. 175.
73 *Ibid.*
74 John McCarthy, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon, « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 », *op. cit.*
75 Minsky et McCarthy créent l'Artificial Intelligence Project au MIT. Puis McCarthy fonde le Stanford AI Laboratory.

> L'étude s'appuie sur la conjecture que tous les aspects de l'apprentissage ou toute autre caractéristique de l'intelligence peuvent en principe être décrits d'une manière si précise qu'une machine peut être fabriquée pour les simuler[76].

Il s'agit bien d'une « conjecture », qui fonde l'étude (« The study is to proceed *on the basis* »). Elle n'est pas prouvée, elle est adoptée en principe. À plusieurs égards, la conjecture se distingue du *credo* de Turing. Il n'est pas directement affirmé que les machines *peuvent* penser (ou que cela peut être prouvé indirectement par un test), mais que « *tous* les aspects de l'apprentissage » peuvent être connus, décrits et simulés[77]. Les auteurs insistent bien sur l'exhaustivité de leur conjecture en précisant que cela concerne « tout autre caractéristique de l'intelligence ». Autrement dit, *aucun* aspect de l'intelligence ne fait exception à la conjecture. La conjecture évoque un niveau de précision sans indiquer *où doit se situer ce niveau.* La description des caractéristiques de l'intelligence doit être *suffisamment précise* pour qu'une machine puisse la simuler. À quel niveau une description est-elle *suffisante* ? S'agit-il de décrire comment une idée en entraîne une autre ou bien les agencements matériels du cerveau ? Au cours de l'été 1956, le séminaire rassemble des chercheurs pour qui cette précision n'a pas le même sens. Certains travaillent sur les réseaux neuronaux (Marvin Minsky, qui change d'avis au cours du séminaire), les machines inductives (Ray Solomonoff) ou le fonctionnement du cerveau (McCulloch fait une brève apparition). D'autres s'intéressent aux mécanismes déductifs – comme John McCarthy, qui souhaite impulser une direction commune au groupe en le faisant travailler sur le jeu d'échecs. D'autres enfin travaillent sur les heuristiques, à l'instar de Newell et Simon qui présentent le Logic Theorist[78]. Une même ambiguïté affecte ce qui concerne le niveau de fabrication de la machine. L'expression anglaise « be made to » (« a machine can be made to simulate it ») est à double sens. Il peut s'agir de « made to » au sens

76 « The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. » *Ibid.* Nous traduisons.

77 Le groupe qui se réunit à Dartmouth n'est qu'indirectement influencé par les travaux de Turing, *via* les chercheurs de la génération cybernétique (en particulier Shannon et McCulloch). D'après Pamela McCorduck « Turing's work had practically no influence on most people at the Dartmouth Conference. For instance, Minsky felt himself much more influenced by McCulloch and Shannon (especially Shannon's early chess paper); Simon considered Turing of no particular influence on his work. » Pamela McCorduck, *Machines Who Think, A Personal Inquiry into the History and Prospects of Artificial Intelligence*, Natick, AK Peters, 2004, p. 113.

78 Un article de Grace Solomonoff donne un aperçu de la teneur des débats et des extraits de notes de session, « Ray Solomonoff and the Dartmouth Summer Research Project in Artificial Intelligence, 1956 ». Sur le même site, on trouve les notes prises par Ray Solomonoff pendant le séminaire http://raysolomonoff.com/dartmouth/dart.html page consultée le 20 août 2020.

de « fabriquée pour » ou bien de « made to » au sens « d'être amenée à ». Dans le premier cas, il s'agit de fabriquer la machine elle-même, dans le deuxième, il s'agit de partir de machines existantes (les ordinateurs) et les « amener à » simuler l'intelligence, c'est-à-dire les programmer. Le premier sens est plus analogique, le deuxième est plus digital.

Malgré cette diversité d'approches, les chercheurs présents à Dartmouth ont pour point commun de croire à la pertinence qu'il y a à entreprendre *en même temps* la fabrication de machines et une enquête sur l'intelligence. Il y a équivalence entre la connaissance (une description « si précise ») et la fabrication. Selon le principe énoncé par Vico, « ce qui est vrai et ce que l'on fait sont convertibles » (« *Verum et factum convertuntur*[79] »). Ils partagent une approche constructiviste de la science : tout comme l'énonçait Turing, c'est la fabrication d'expériences, en l'occurence les machines, qui serviront d'arguments. Bien que les discussions du séminaire abordent de front la perspective de « machines pensantes » (« thinking machines[80] »), le texte de la proposition est plus prudent. Il se contente d'évoquer la « simulation » des « caractéristiques de l'intelligence », ce qui laisse flotter un doute quant au fait que l'objectif serait de fabriquer des machines qui *pensent effectivement* ou bien des machines qui, sans être intelligentes, sont capables de « simuler » les caractéristiques de l'intelligence. Le texte de la proposition fait principalement référence à l'« intelligence », plutôt qu'à la « pensée ». Il y a exception au paragraphe sept (« Randomness and creativity »), où est évoquée la différence entre « la pensée créative » et « la pensée compétente sans imagination », ce qui amène les auteurs à évoquer l'intuition : cette dernière doit « guider » une certaine dose de hasard pour permettre une « pensée créative » dans ce qui est sinon « une pensée ordonnée ». Nous commenterons ce paragraphe ultérieurement. Pour le moment, remarquons simplement que les auteurs ne semblent *pas faire de distinction entre l'intelligence et la pensée.* Cela est confirmé par les notes prises pendant le séminaire par Solomonoff[81] montrant que de nombreuses discussions ont évoqué les « machines pensantes », alors que les participants se sont mis d'accord pour baptiser leur projet « intelligence artificielle ».

79 Vico le formule dès 1710 dans un de ses premiers ouvrages, *De antiquissima italorum sapienta*. Jean-Pierre Dupuy commente l'adoption de ce principe par les sciences cognitives dans *Aux origines des sciences cognitives,* Paris, La Découverte, 2005, p. 16.

80 Les notes de Ray Solomonoff sont intitulées T.M. pour « thinking machines » et évaluent les différentes pistes pour arriver à une « machine pensante ». Ainsi, dans les pages de conclusion (« Overall summary »), rédigées le 18 août 1956, il remarque que « these guys may eventually invent a T.M. , simply by working more and more interesting special problems. Simon and Newell ; Minksy : best candidates – Trench[ard] More is a question mark. »

81 Celles-ci peuvent être consultées sur le site http://raysolomonoff.com/dartmouth/dart.html, page consultée le 20 août 2020.

1.1.8. Nommer l'intelligence artificielle

Lorsqu'à l'été 1956 des chercheurs d'horizons divers se réunissent pour travailler ensemble à un objet commun, le nom de ce dernier n'est pas arrêté. Pendant le séminaire, Ray Solomonoff reçoit ainsi une lettre adressée au « Symposium sur l'intelligence de synthèse » (« Symposium on Synthetic Intelligence »). Certains évoquent les machines pensantes, d'autres préfèrent s'en tenir à des théories existantes, comme la cybernétique ou la théorie des automates[82]. Le terme choisi par McCarthy au moment de la préparation du séminaire – intelligence artificielle –, soulève des réticences[83]. Arthur Samuel trouve que « le mot artificiel laisse penser qu'il y a quelque chose de bidon [phony] là-dedans », « que c'est entièrement artificiel et qu'il n'y a rien de réel du tout dans ce travail[84]. » Newell et Simon, qui ont effectué les recherches les plus abouties (ce sont les seuls du groupe à présenter un programme « intelligent » en état de marche), préfèrent parler de « complex information processing ». Mais il faut un terme assez large pour inclure la notion d'intelligence et la perspective de machines intelligentes. Peu de temps auparavant, pour un ouvrage collectif réalisé avec Claude Shannon, ce dernier impose à McCarthy le terme « sobre et scientifique » de *Automata Studies*[85], pour éviter un nom « trop *flashy*[86] ». À la déception de McCarthy, les auteurs prennent le nom au sens strict et envoient des articles spécialisés sur les « principes mathématiques sous-jacents aux systèmes électromécaniques » au lieu d'aborder « la relation entre langage et intelligence » ou « la possibilité pour les machines de jouer à des jeux[87] ». La formule de McCarthy a le mérite d'éviter un nouveau malentendu et de permettre que soit effectivement débattu le sujet des machines intelligentes. D'autre part, si l'expression est « trop flashy », elle participe, selon les

82 « Some other names were cybernetics, automata theory, complex information processing, [16, p. 115] or just "thinking machines." » Grace Solomonoff, *op. cit.*

83 « Although the conference was officially called The Dartmouth Summer Research Project on Artificial Intelligence, many attendees balked at that term, invented by McCarthy », Pamela McCorduck, *op. cit.*, p. 114.

84 « "The word artificial makes you think there's something kind of phony about this," says Arthur Samuel, "or else it sounds like it's all artificial and there's nothing real about this work at all." », Pamela McCorduck, *op. cit.*, p. 115.

85 Claude Shannon et John McCarthy (ed.), *Automata Studies*, Princeton, Princeton University Press / « Annals of Mathematics Studies », n°34, 1956.

86 « McCarthy wanted to use a term different from automata studies for the papers he hoped to get for the book, but Shannon objected that any other phrase was simply too flashy, that the theory of automata would be sober and scientific. » Pamela McCorduck, *op. cit.*, p. 115.

87 « Most of the papers they received for the book were in fact about automata theory in the narrowest sense, that is, mathematical principles underlying the operation of electromechanical systems, and not about the relation of language to intelligence, or the ability of machines to play games, or any of the other topics McCarthy was becoming more and more fascinated by. » *Ibid.*

termes de Ganascia, à « frapper les esprits[88] », ce qui s'avèrera crucial au moment de lever des fonds. Après la conférence de Dartmouth, le terme rencontre un succès académique et médiatique grâce aux efforts de Marvin Minsky et John McCarthy. C'est ce terme qui sert à désigner le groupe de recherche qui se constitue au MIT (qui deviendra le Computer Science and Artificial Intelligence Laboratory ou CSAIL), ainsi que le laboratoire fondé par McCarthy à Stanford (Stanford AI Laboratory).

Newell et Simon continuent pendant des années à désigner leur travail sous les termes de « complex information processing[89] ». Mais vingt-quatre ans plus tard, lorsque les chercheurs se regroupent en association, c'est l'expression de McCarthy qui prévaut. L'association est baptisée « Association for the Advancement of Artificial Intelligence » (AAAI). Nommé président, Newell abandonne ses premières réticences et prend acte du consensus qui s'est installé : « chérissez le nom d'intelligence artificielle. C'est un bon nom. Comme tous les noms de champs scientifiques, il grandira jusqu'à devenir exactement ce que son champ signifie[90]. » Avec cette formule alambiquée (« it will grow to become exactly what its field comes to mean »), Newell propose d'embrasser le nom d'intelligence artificielle *pour la même raison* qui l'avait amené à le refuser : le champ scientifique n'est pas à la hauteur de son nom puisque les programmes fabriqués ne sont pas intelligents. Mais au lieu de l'interpréter comme un mensonge, Newell choisit d'*en faire une promesse. «* Intelligence artificielle » ne désigne pas les machines d'aujourd'hui mais ce qu'elles doivent devenir. La pertinence du nom est renvoyée à l'avenir, tout comme Turing renvoyait à l'avenir la pertinence de son argument : le nom deviendra juste, et Turing finira par avoir raison, une fois qu'auront été fabriquées des machines intelligentes (ou pouvant passer pour intelligentes). En souscrivant progressivement à ce nom, la communauté des chercheurs s'accorde sur cette promesse et adopte comme horizon de recherche la réalisation de machines intelligentes.

88 Jean-Gabriel Ganascia, *L'intelligence artificielle,* Éditions Le Cavalier Bleu, 2007, p. 8.

89 « Neither Newell nor Simon liked the phrase, and called their own work complex information processing for years thereafter. But artificial intelligence is the phrase that stuck. » Pamela McCorduck, *op. cit.*, p. 115.

90 « So cherish the name artificial intelligence. It is a good name. Like all names of scientific fields, it will grow to become exactly what its field comes to mean. » Allen Newell, « The First AAAI President's Message », *AI Magazine,* Winter 2005, 25th anniversary issue. Le fait est remarqué par Nils Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements,* Cambridge University Press, 2009, p. 79.

1.1.9. En sciences cognitives, il n'y a pas d'*ignorabimus*

La conjecture formulée à l'occasion du séminaire de Dartmouth pourrait sembler anodine – quoi de plus normal qu'une science naissante fasse l'hypothèse que l'objet qu'elle se donne puisse être décrit par les moyens qu'elle s'invente ? – s'il n'y avait l'insistance sur l'exhaustivité : ce sont *« tous* les aspects de l'apprentissage » et *« toute autre caractéristique* de l'intelligence » qui peuvent être ainsi « décrits ».

Tout comme Hilbert affirmait qu'« en mathématiques, il n'y a pas d*'ignorabimus,* nous pouvons toujours trouver une réponse à une question pourvu qu'elle ait un sens[91] », les fondateurs de l'intelligence artificielle comme champ scientifique énoncent dès le premier paragraphe de leur texte princeps qu'en sciences cognitives, il n'y a pas *d'ignorabimus* : il est toujours possible de décrire une caractéristique de l'intelligence d'une manière si précise qu'une machine puisse être fabriquée pour la simuler. Autrement dit, *aucun* aspect de l'intelligence ne résiste à la description. De la même manière que les théorèmes d'incomplétude ont porté un coup au programme de Hilbert, de nombreux auteurs, inspirés par les travaux de Gödel[92], et à commencer par Gödel lui-même[93], se sont mis en quête d'un « résultat négatif » qui invalide l'ambition du projet d'intelligence artificielle. Mais si la portée des théorèmes d'incomplétude est reconnue dans le champ des mathématiques, la pertinence de leur extrapolation à l'intelligence artificielle fait toujours débat[94]. Ils laissent planer un doute sur la pertinence des hypothèses du projet d'intelligence artificielle, mais n'ont pas permis de produire de « résultat négatif » aussi incontestable – et peut-être en sont-ils incapables[95].

---

91 David Hilbert, « Problèmes de fondation des mathématiques », *in* Jean Largeault, *Intuitionnisme et théorie de la démonstration*, Paris, Vrin, 1992, p. 185. Le terme *ignorabimus* est une référence à la phrase d'Emil du Bois-Reymond, « *ignoramus et ignorabimus* », c'est-à-dire « nous ignorons et nous ignorerons ». Nous revenons ultérieurement sur cette référence.

92 En particulier John Lucas, « Minds, Machines, and Gödel », *Philosophy,* n° 36, 1961, p. 112-127, repris ensuite par Roger Penrose, *Emperor's New Mind*, Oxford, Oxford University Press, 1989.

93 Gödel ne critique pas directement la conjecture formulée à Dartmouth mais discute la thèse de Turing au sujet de l'existence de procédure finies non mécaniques. De telles procédures pourraient être considérées comme des « aspects » de l'intelligence non descriptibles « de manière si précise qu'une machine peut les simuler ». Voir Pierre Cassou-Noguès, « Gödel et la thèse de Turing », *Revue d'histoire des mathématiques,* n°14, 2008, p. 77-111, et Inês Hipolito, « Gödel on the mathematician's mind and Turing Machines », *E-Logos Electronic Journal for Philosophy*, n°22, 2014.

94 Les arguments de Lucas et Penrose ont été critiqués par Putnam, Benacerraf, Quine et bien d'autres. Voir Stanislaw Krajewski, « On Gödel Theorem and Mechanism : Inconsistency or Unsoundness is Unavoidable in Any Attempt to 'Out-Gödel' the Mechanist », *Fundamenta Informatica*, n°81, 2007, p. 1-9.

95 Voir section 2.5.3. Les limites du formalisme au secours de l'intuition.

1.1.10. Le paradoxe du contre-argument

Pour invalider la conjecture selon laquelle « tous les aspects de l'apprentissage ou tout autre caractéristique de l'intelligence » peuvent être décrits, il faudrait présenter au moins une « caractéristique de l'intelligence » *qui ne puisse pas faire l'objet d'une description* assez précise[96]. Tout comme la seconde loi de la thermodynamie a cloué le bec aux tenants du mouvement perpétuel, on aimerait pouvoir exhiber une caractéristique qui constituerait une objection indéniable. Mais l'entreprise n'est pas sans paradoxe : comment la présenter s'il n'est pas possible de la décrire ? Cela a-t-il du sens d'évoquer une « caractéristique » de l'intelligence dont nous saurions qu'elle existe sans pour autant être capable de la décrire ? Voici un petit dialogue illustrant l'aporie à laquelle cela mène :

> *Les « anti-IA » :* Le projet d'intelligence artificielle n'est pas possible car nous avons la faculté de nous émouvoir (ou bien : de créer, de choisir, de comprendre, etc.). Cette faculté n'est pas réplicable, elle est un privilège de l'humain (ou du vivant).
>
> *Les « pro-IA » :* L'émotion (ou la créativité, la décision, la compréhension, etc.) sont des notions floues que je ne peux accepter comme une objection valable. Pourriez-vous, d'une part, me décrire cette faculté avec précision, et d'autre part, me prouver qu'elle est nécessaire à l'intelligence ? Pourquoi ne pourrait-il y avoir d'intelligence sans émotion ?
>
> *Les « anti-IA » :* Nous pouvons vous fournir des travaux montrant que les émotions sont indispensables au raisonnement[97], ainsi que des travaux décrivant les principales émotions et leur fonctionnement[98]. Vous ne pourrez pas nous accuser de rester vague !
>
> *Les « pro-IA » :* Très bien. Dans ce cas nous pouvons intégrer votre description dans un modèle informatique et demander à nos robots d'exprimer les émotions mentionnées[99].

96 Nous mettons de côté les considérations sur le *niveau de précision requis* pour que la description permette une simulation.

97 Voir par exemple les ouvrages d'Antonio Damasio.

98 Par exemple, les travaux de Robert Plutchik.

99 A titre d'exemple, Matthieu Courgeon. *Marc : modèles informatiques des émotions et de leurs expressions faciales pourl'interaction Homme-machine affective temps réel.* Thèse de doctorat en Intelligence artificielle [cs.AI] soutenue à l'Université Paris Sud- Paris XI, 2011.

Tout un sous-domaine de l'intelligence artificielle et de la robotique va s'y consacrer : l'« affective computing ».

Les « anti-IA » : Non ce n'est pas cela. Vos modèles sont trop simplistes. L'émotion, c'est bien plus que cela.

Les « pro-IA » : Ah bon, mais de quoi s'agit-il alors ? Que faut-il ajouter ? Ou bien vous nous en donnez une description précise (et nous nous empresserons de l'ajouter à notre modèle), ou bien votre objection n'est pas recevable.

Pro et anti IA (ensembles) : Nous finirons bien par avoir le dernier mot !

La conjecture de Dartmouth telle qu'elle est formulée exclut la possibilité d'une invalidation par un contre-exemple clair. Si le contre-exemple est bien descriptible, il peut être simulé par une machine, et ce n'est pas un contre-exemple. Si, à l'inverse, le contre-exemple est peu descriptible, il ne sera pas reçu comme un contre-exemple sérieux. On lui reprochera d'être flou et mal défini. Cela rend irrecevable, pour les tenants du projet d'intelligence artificielle, les approches qui expliquent l'intelligence par le vague[100] et l'absence d'organisation du cerveau, voire sa capacité à l'erreur[101], comme étant les clefs de l'intelligence.

Il est facile de balayer les différentes « facultés » présentées en objection au projet d'intelligence artificielle – comme les émotions. Ou bien on dira que les émotions sont une notion trop vague pour être prise en considération. Ou bien on donne une description simpliste mais claire d'un « mécanisme des émotions » et dans ce cas rien n'empêche d'intégrer le « mécanisme » en question à un projet d'« affective computing ». Plus généralement, s'il est facile d'admettre que nous ignorons une bonne partie du fonctionnement de notre pensée, comment établir quoi que ce soit au sujet de cette part ignorée ? Comment établir si elle peut ou non faire l'objet d'une description ? Comment évoquer rationnellement un angle mort si à aucun moment il n'est possible de l'éclairer ?

100 « The very vagueness [of the cerebral hemispheres] constitutes their advantage… An organ swayed by slight impressions is an organ whose natural state is one of unstable equilibrium » William James, cité par David Bates, « Automaticity, Plasticity, and the Deviant Origins of Artificial Intelligence », in Bates et Bassiri (eds.), *Plasticity and Pathology : On the Formation of the Neural Subject,* New York, Fordham University Press, 2015, p. 201.

101 Pour Pierce « Intelligence was, in a sense considered to be a consequence of a certain disorganization and unpredictability, and potentially even pathological disorder might explain the leaps of a genius intelligence. » David Bates, « Automaticity, Plasticity, and the Deviant Origins of Artificial Intelligence », *op. cit.*, p. 202.

1.1.11. L'intuition est-elle l'*ignorabimus* des sciences cognitives ?

Dans sa thèse soutenue en 1938, Turing défend l'existence de l'intuition, qu'il oppose au raisonnement explicite : « l'activité de l'intuition consiste à opérer des jugements spontanés qui ne sont pas le résultat d'une suite consciente de raisonnements[102] ». Elle est d'une importance primordiale pour le mathématicien, puisque c'est elle qui apporte les propositions que le raisonnement explicite se charge ensuite de prouver[103]. Et elle intervient également, ainsi qu'il le précise un an plus tard dans une lettre à Max Newman, dans le choix du raisonnement explicite à utiliser[104]. Aussi, « l'impossibilité de trouver une logique formelle qui élimine entièrement la nécessité d'utiliser l'intuition », amène Turing à considérer « un système logique dans lequel toutes les étapes ne sont pas mécaniques, certaines étant intuitives[105] », c'est-à-dire à faire *coexister* le raisonnement explicite et l'intuition. Bien qu'il défende l'existence, l'importance, et la nécessité de l'intuition, Turing *renonce à en faire une description précise* : « Je n'essayerai pas d'expliquer cette idée 'd'intuition' de manière plus explicite[106]. » La

102 « The activity of the intuition consists in making spontaneous judgments which are not the result of conscious trains of reasoning. » Alan Turing, « Systems of Logic Based on Ordinals », in Jack 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, 2004, p. 192.

103 Ainsi le raisonnement explicite, qu'il appelle « ingenuity » sert à vérifier les propositions de l'intuition : « The exercise of ingenuity in mathematics consists in aiding the intuition through suitable arrangements of propositions, and perhaps geometrical figures or drawings. » Les propositions de l'intuition n'étant pas toujours vraies « These judgments are often but by no means invariably correct […]. » *Ibid.*

104 « The choice of a proof checking machine involves intuition […] » Alan Turing, « Letters on Logic to Max Newman (c. 1940) », in Jack Copeland, *op. cit.*, p. 215.

105 Nous restituons l'intégralité du paragraphe, Turing y résumant sa démarche : « In pre-Gödel times it was thought by some that it would probably be possible to carry this programme to such a point that all the intuitive judgments of mathematics could be replaced by a finite number of these rules. The necessity for intuition would then be entirely eliminated. In our discussions, however, we have gone to the opposite extreme and eliminated not intuition but ingenuity, and this in spite of the fact that our aim has been in much the same direction. We have been trying to see how far it is possible to eliminate intuition, and leave only ingenuity. We do not mind how much ingenuity is required, and therefore assume it to be available in unlimited supply. In our metamathematical discussions we actually express this assumption rather differently. We are always able to obtain from the rules of a formal logic a method of enumerating the propositions proved by its means. We then imagine that all proofs take the form of a search through this enumeration for the theorem for which a proof is desired. In this way ingenuity is replaced by patience. In these heuristic discussions, however, it is better not to make this reduction. In consequence of the impossibility of finding a formal logic which wholly eliminates the necessity of using intuition, we naturally turn to ''non-constructive'' systems of logic with which not all the steps in a proof are mechanical, some being intuitive. » Alan Turing, « Systems of Logic Based on Ordinals », *op. cit.*, p. 192-193.

106 « I shall not attempt to explain this idea of ''intuition'' any more explicitly. » *Ibid.*, p. 192.

mention de l'intuition renvoie à une ignorance[107]. De la même manière, il s'abstient d'entrer dans les détails de la « machine oracle » censée intégrer l'intuition au raisonnement du mathématicien : « Nous n'en dirons pas plus quant à la nature de cet oracle sinon qu'il ne peut être une machine[108] ». En tant que « caractéristique de l'intelligence » du mathématicien qui « ne peut être une machine », l'« oracle » de Turing pourrait fournir la matière d'une objection au projet d'intelligence artificielle, si seulement il en donnait une définition plus précise. Mais il n'en propose qu'une définition négative, en opposition à la machine et au raisonnement explicite, sans lui attribuer de caractéristique propre.

Ainsi, bien que les limites de la pensée formelle laissent planer un doute quant à la pertinence du projet d'intelligence artificielle, doute renforcé par l'expérience des mathématiciens témoignant de leur usage de l'intuition et de la nécessité de lui faire une place[109], cela ne suffit pas à invalider les hypothèses formulées à Dartmouth dans la mesure où manque une définition de l'intuition, ainsi qu'une preuve de son existence. Tout se passe comme si nous ne pouvions faire mieux que de *partager l'intuition de l'existence de l'intuition* sans que cette intuition ne puisse trouver de confirmation par le raisonnement explicite. Dès lors, la pertinence de l'hypothèse formulée à Dartmouth est laissée en suspens, puisque l'intuition n'est ni abandonnée (ce qui validerait l'hypothèse), ni définie ou fondée (ce qui l'invaliderait). Cette situation laisse le champ libre aux chercheurs en intelligence artificielle, qui se sentent en droit de considérer que la fabrication d'une intelligence « générale » constitue le meilleur moyen de confirmer leurs hypothèses.

Si le programme d'intelligence artificielle a vécu un changement de paradigme[110] récent, il repose toujours sur la conjecture de Dartmouth[111]. Grâce au renouveau des réseaux de neurones et à l'invention des algorithmes d'apprentissage profond, les chercheurs pensent être

107 Pour Jean Lassègue, ce sont « les deux facultés d'intuition et d'ingéniosité » qui « ne sont pas accessibles dans leur intégralité : dans le cas de l'intuition, c'est la capacité de *production* qui reste inaccessible globalement du fait que l'on ignore la 'chaîne de raisonnement' qui rendrait compte de ses manifestations. Dans le cas de l'ingéniosité en revanche, c'est notre capacité d'anticipation du comportement de la machine universelle qui reste inconnue dans sa globalité. » Jean Lassègue « L'évolution du constructivisme turingien : de la logique à la morphogenèse », *Intellectica. Revue de l'Association pour la Recherche Cognitive*, n°39, 2004/2, p. 112.

108 « We shall not go any further into the nature of this oracle apart from saying that it cannot be a machine. » *Ibid.*, p. 165.

109 Nous revenons ultérieurement sur l'expérience de l'intuition chez les mathématiciens. Voir chapitre 2.5 Peut-on se passer de l'intuition ?

110 Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones, L'invention des machines inductives et la controverse de l'intelligence artificielle », in *Réseaux,* n°211, 2018 / 5, p. 173-220. Nous détaillons ce changement de paradigme dans le chapitre suivant.

111 « En dépit des progrès époustouflants enregistrés ces dernières années, l'étude de l'intelligence artificielle repose toujours sur la même conjecture que rien, jusqu'à présent, n'a permis ni de démentir, ni de démontrer irréfutablement. » Jean-Gabriel Ganascia, *L'intelligence artificielle, Vers une domination programmée ?* Paris, Le Cavalier Bleu, 2017.

en passe de doter les machines d'intuition, et de prouver que l'intuition est, comme toutes les « caractéristiques de l'intelligence », descriptible d'une manière « si précise qu'une machine peut être fabriquée pour [la] simuler ». Ils seraient sur le point d'invalider les propos tenus par Turing en 1938 (l'intuition « ne peut être une machine »), pour confirmer ceux tenus en 1950 : la possibilité de fabriquer une machine pensante. Afin de pouvoir examiner ces prétentions, nous commencerons par décrire l'objet d'une si grande ambition : les réseaux de neurones.

1.2. Histoire des réseaux de neurones

1.2.1. Une place à part

Bien que les réseaux de neurones ne soient qu'une des très nombreuses méthodes inventées par les chercheurs en intelligence artificielle, ils occupent une place particulière dans l'histoire de ce champ. Leur invention par McCulloch et Pitts est la première tentative de décrire l'opération des neurones (et la pensée) comme une machine de Turing. Si bien que certains considèrent l'article de 1943 comme l'acte de naissance du projet d'intelligence artificielle, bien avant la conférence de Dartmouth[112]. Dans les années 1950, la recherche en intelligence artificielle prend des directions très variées, mais les réseaux de neurones occupent le devant de la scène. Marvin Minsky, qui deviendra un champion du paradigme symbolique, fait ses premiers travaux sur les réseaux de neurones[113]. Le Perceptron de Frank Rosenblatt, inventé en 1957, déclenche un enthousiasme démesuré auprès de la presse et du grand public[114]. En 1968, une controverse à son sujet provoque le premier « hiver de l'IA[115]». Les réseaux de neurones pâtissent de cette controverse et sont mis au ban de la recherche pour quelques dizaines d'années, pendant lesquelles l'approche symbolique tient le haut du pavé. Aujourd'hui, avec l'invention de l'algorithme de *back-propagation* et du *deep learning,* les réseaux de neurones sont revenus sur le devant de la scène et ont ouvert un nouveau chapitre de l'histoire de l'intelligence artificielle. Ils sont les candidats privilégiés des chercheurs qui évoquent la perspective de doter les machines d'intuition.

112 C'est le choix que fait Margaret Boden dans *The Philosophy of Artificial Intelligence,* Oxford, Oxford Readings in Philosophy, 1990. Cette relecture de l'histoire, fréquente aujourd'hui, signale aussi la « victoire » du paradigme connexionniste sur le paradigme symbolique.

113 Le « Stochastic Neural Analog Reinforcement Calculator » ou SNARC. Voir Daniel Crevier, *A la recherche de l'intelligence artificielle,* Paris, Champs / Flammarion, 1997.

114 Une citation du *New York Times* au sujet du Perceptron est devenue un classique des exagérations au sujet de l'IA. Pour le journaliste, le Perceptron est « the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence. » voir Mikel Olazaran, « A Sociological Study of the Official History of the Perceptrons Controversy » *Social Studies of Science*, vol. 26, No. 3, Août 1996, p. 611-659.

115 Avec le livre *Perceptrons* de Minsky et Papert, voir Mikel Olazaran, *op. cit.*

1.2.2. McCulloch et Pitts, l'invention des réseaux de neurones (1943)

La notion de « réseau de neurones » naît de la collaboration entre McCulloch, neurophysiologiste établi à l'Université de Chicago, dont l'ambition affirmée est de comprendre comment fonctionne le cerveau pour expliquer la pensée, et du jeune Walter Pitts qui met ses facilités en mathématiques et sa connaissance des langages formels au service de l'ambition de son aîné[116]. En 1943, ils publient ensemble l'article « A Logical Calculus of the Ideas Immanent in Nervous Activity[117] » qui vise à montrer que la structure du cerveau, telle qu'on la connaît, peut être vue comme effectuant du calcul propositionnel. Les auteurs retiennent pour caractéristique principale du neurone le fait qu'il envoie, ou non, une impulsion électrique en fonction des impulsions reçues par d'autres neurones. À l'instar des propositions de l'algèbre booléenne, les neurones ont un aspect binaire : ils transmettent ou ne transmettent pas d'impulsion électrique, c'est « tout ou rien » (« all-or-none »). On peut donc comparer l'état d'un neurone à une proposition qui sera vraie ou fausse (transmission ou absence de transmission[118]) en fonction de conditions préalables – les impulsions reçues des neurones précédents – eux aussi équivalents à des propositions affectées de valeurs de vérité. Au niveau de chaque neurone il y a 'computation' des impulsions reçues (des valeurs de vérité) des neurones précédents qui amène à décider d'envoyer ou non une impulsion (donner une valeur de vérité à la proposition[119]). Un réseau de neurones est donc comme un dispositif de calcul propositionnel au sens où il permet d'opérer des inférences (« si ceci, alors cela »). L'idée de matérialiser la logique booléenne via un circuit électrique avait déjà été élaborée par Claude Shannon dans son mémoire de fin d'études[120]. Sans mentionner les travaux de Shannon,

116 Gualtiero Piccinini détaille le contexte de l'article et en propose un commentaire dans « The First Computational Theory of Mind and Brain : A Close Look at McCulloch and Pitts's 'Logical Calculus of Ideas Imminent in Nervous Activity », *Synthese*, 141 : 175–215, 2004.

117 McCulloch, Warren, et Walter Pitts, « A Logical Calculus of the Ideas Immanent in Nervous Activity ». *The Bulletin of Mathematical Biophysics*, vol. 5, no 4, décembre 1943, p.115-33.

118 « The 'all-or-none' law of nervous activity is sufficient to insure that the activity of any neuron may be represented as a proposition. », *Ibid.*, p. 117.

119 « To each reaction of any neuron there is a corresponding assertion of a simple proposition. This, in turn, implies either some other simple proposition or the disjunction or the conjunction, with or without negation, of similar propositions, according to the configuration of the synapses upon and the threshold of the neuron in question. » *Ibid.*

120 Claude Shannon, « A Symbolic Analysis of Relay and Switching Circuits », *Transactions of the American Institute of Electrical Engineers*, Volume: 57, Issue: 12, 1938. Bien avant Shannon, l'idée avait été formulée, au détour d'une lettre, par Pierce, « Letter to A. Marquand, December 30, 1886 », in C. Kloesel et al. (eds.), *Writings of Charles S. Pierce : A Chronological Edition* (Bloomington, Indiana University Press, 1993), 5 : 421-22, cité par David Bates, « Automaticity, Plasticity, and the Deviant Origins of Artificial Intelligence », in Bates et Bassiri (eds.), *Plasticity and Pathology : On the Formation of the Neural Subject*, New York, Fordham University Press, 2015.

McCulloch et Pitts font un pas supplémentaire en assimilant le circuit au cerveau. Ils décrivent plusieurs exemples de réseaux de neurones et détaillent leur fonctionnement en recourant à un langage formel emprunté aux *Principia Mathematica* de Russell et Whitehead, ainsi qu'à Carnap. En conclusion, ils font une comparaison entre les réseaux de neurones tels qu'ils les ont décrits et une machine de Turing (« each of the latter numbers [les nombres que peut calculer une machine de Turing] can be computed by such a net[121] »), et laissent entendre que les réseaux de neurones sont un dispositif capable de calculer tout nombre calculable, autrement dit d'effectuer n'importe quel algorithme, à l'instar d'une machine de Turing[122]. Pour McCulloch et Pitts, les réseaux de neurones ne sont pas un modèle simplifié du cerveau, une métaphore, ou une hypothèse ambitieuse à vérifier. Le cerveau fait *réellement* du calcul propositionnel au niveau des neurones. Cela pourrait suffire à expliquer nos idées, comme l'indique le titre de leur article (le « calcul logique » décrit est celui des « idées immanentes à l'activité nerveuse »), et l'ensemble de nos états mentaux : « Ainsi, aussi bien l'aspect formel que final de cette activité que nous appelons mentale sont rigoureusement déductibles de l'état actuel de la neurophysiologie[123] ». Et ils se réjouissent de ce que cela implique pour le traitement de la folie :

> Certainly for the psychiatrist it is more to the point that in such systems "Mind" no longer "goes more ghostly than a ghost."[124] Instead, diseased mentality can be understood without

121 « One more thing is to be remarked in conclusion. It is easily shown: first, that every net, if furnished with a tape, scanners connected to afferents, and suitable efferents to perform the necessary motor-operations, can compute only such numbers as can a Turing machine; second, that each of the latter numbers can be computed by such a net; and that nets with circles can be computed by such a net; and that nets with circles can compute, without scanners and a tape, some of the numbers the machine can, but no others, and not all of them. This is of interest as affording a psychological justification of the Turing definition of computability and its equivalents, Church's definability and Kleene's primitive recursiveness: If any number can be computed by an organism, it is computable by these definitions, and conversely. » McCulloch et Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity », *op. cit.*, 129.

122 C'est une revendication exagérée. Piccinini en donne commentaire : « But in discussing computation in their paper, McCulloch and Pitts did not prove any results about the computation power of their nets; they only stated that there were results to prove. And their conjecture was not that their nets can compute anything that can be computed by Turing Machines. Rather, they claimed that if their nets were provided with a tape, scanners, and ''efferents,'' then they would compute what Turing Machines could compute; without a tape, McCulloch and Pitts expected even nets with circles to compute a smaller class of functions than the class computable by Turing Machines. » Gualtiero Piccinini, *op. cit.*, p. 198. L'étude du modèle formel de McCulloch et Pitts, et de ce qu'il peut calculer, est repris ultérieurement par Kleene dans un article écrit en 1951 pour la RAND et publié en 1956 dans l'ouvrage collectif déjà mentionné *Automata Studies, op. cit.*

123 « Thus both the formal and the final aspects of that activity which we are wont to call mental are rigorously deduceable from present neurophysiology. » McCulloch et Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity », *op. cit.*

124 L'expression fait référence aux travaux du neurophysiologiste Sherrington et au fait que l'esprit échappe à l'oeil du physiologiste. L'idée d'un « fantôme dans la machine » est commentée ultérieurement par Gilbert Ryle dans son ouvrage de 1949. *La notion d'esprit : pour une critique des concepts mentaux*, Paris, Payot, 1978. Vulgarisée grâce à l'essai d'Arthur Koestler, *Ghost in the machine* (1967), la notion devient un *topos*

> loss of scope or rigor, in the scientific terms of neurophysiology. For neurology, the theory sharpens the distinction between nets necessary or merely sufficient for given activities, and so clarifies the relations of disturbed structure to disturbed function[125].

Le traitement de la folie pourrait dépendre de l'étude de la matière (la biologie du cerveau) et des mathématiques (les modèles de neurones), c'est-à-dire dépendre de sciences « dures » et ainsi bénéficier de leur rigueur. Avec l'ambition de décrire le fonctionnement du cerveau et de la pensée vient celle de définir et prescrire ce que c'est que de *bien penser.* La recherche d'un modèle du cerveau déborde vers la promotion d'un cerveau modèle.

1.2.3. Premier âge d'or pour les réseaux de neurones (1951-1969)

Après la guerre, les réseaux de neurones formels de McCulloch et Pitts inspirent un certain nombre d'algorithmes. En 1951, Marvin Minksy invente le 'Stochastic Neural Analog Reinforcement Calculator', ou SNARC. En 1955, lors d'une conférence à Los Angeles, Clark et Farley présentent leurs travaux sur l'application des réseaux de neurones à la reconnaissance de formes ('pattern recognition'). En 1957, Frank Rosenblatt propose le Perceptron, un réseau de neurones déployé sur une machine spécialisée, le Mark I. Il est capable de reconnaître les lettres de l'alphabet (à condition de ne pas changer la police de caractère) et suscite un enthousiasme immense, largement nourri par les déclarations enflammées de Rosenblatt.

> En 1958, Rosenblatt le présenta à la presse avec cette emphase caractéristique qui allait finalement lui valoir les foudres de la communauté scientifique. Dans un article intitulé « Les cerveaux humains remplacés ? » le prestigieux magazine *Science* décrivit l'invention comme n'étant « pas un esprit mécanique ordinaire qui emmagasine puis régurgite l'information […]. Le Perceptron sera finalement capable d'apprendre, de prendre des décisions et de traduire des langues[126]. »

Grâce au Perceptron de Rosenblatt, mais également aux réseaux de Taylor, aux algorithmes Adaline et Madaline de Widrow et Hoff, aux 'matrices d'apprentissage' de Karl Steinbuch, « le

---

des paradoxes du dualisme corps-esprit, jusqu'à inspirer le film d'animation de Mamoru Oshii, *Ghost in the shell* (1995).

125 McCulloch et Pitts, *op. cit.*, p. 132.

126 Daniel Crevier, *A la recherche de l'intelligence artificielle*, Paris, Flammarion, 1997, p. 127.

domaine des réseaux neuronaux artificiels conn[ait] une première période d'épanouissement quasiment tout au long des années 1960[127]. »

1.2.4. Comment « apprend » un réseau de neurones ?

Un Perceptron de Rosenblatt est composé d'« unités sensorielles » (quatre cents cellules photoélectriques), suivies d'« unités d'association » (une couche de neurones) et d'une « unité d'activation » (une lampe qui s'allume). En fonction de la forme des objets présentés au Perceptron, certaines de ses « unités sensorielles » (les capteurs photoélectriques) s'activent et envoient un signal aux « unités d'association » ou « neurones ». Ces derniers réalisent trois opérations :

- additionner les signaux reçus depuis les différents capteurs. Chaque signal est multiplié par un « poids » renforçant ou diminuant sa prise en compte ;
- la somme pondérée des signaux reçus est comparée avec un seuil ;
- si ce seuil est dépassé, le neurone envoie un signal à l'« unité d'activation » : une lumière s'allume, annonçant qu'une forme spécifique a été « reconnue ».

Autrement dit, le dispositif du Perceptron revient à associer la présence de certaines combinaisons d'activation des cellules photoélectriques au déclenchement de la lumière. Pour peu que les « poids » soient bien attribués, le Perceptron est capable de signaler la présence de formes visuelles élémentaires. L'invention de Rosenblatt permet de « percevoir » une variété de formes en combinant plusieurs Perceptrons, chacun étant calibré pour une forme donnée.

Les poids sont représentés par des potentiomètres et modulés à l'aide de petits moteurs électriques. Ils sont d'abord fixés de manière aléatoire puis ajustés au fur et à mesure. Ils sont diminués si la lumière s'allume à tort, et augmentés si la lumière ne s'allume pas quand il le faut, jusqu'à ce que la lumière s'allume au bon moment[128]. Ainsi, l'« apprentissage[129] », c'est-à-dire la recherche de la bonne combinaison de poids permettant de reconnaître une forme

127 *Ibid.,* p. 130.

128 « […] les connexions de départ étaient fixées de façon aléatoire, Rosenblatt voulant simuler par là les connexions neuronales du cerveau d'un petit enfant. Au début de la période de formation, n'importe quelle lumière ou série de lumières pouvait donc s'allumer. S'il ne sagissait pas de la bonne, Rosenblatt diminuait alors d'une quantité fixe les coefficients de poids qui influaient sur ses connexions avec les unités sensorielles, et au contraire les augmentait lorsqu'une lumière qui aurait dû s'allumer ne s'allumait pas. Lorsque la bonne lumière s'allumait, il ne modifiait rien. » Daniel Crevier, *Ibid.,* p. 128-129.

129 Le procédé est inspiré du principe formulé par le neuropsychologue Donald Hebb : les poids synaptiques se renforcent quand les neurones doivent être actifs ou inactifs simultanément, et se réduisent quand les neurones doivent avoir des comportements opposés.

donnée, se fait par essai-erreur, par tâtonnement. On peut comparer la recherche de la bonne combinaison de poids comme un jeu de 'chaud-froid' : un joueur à qui on a bandé les yeux cherche un objet dans une pièce, guidé par un joueur voyant qui lui indique s'il 'chauffe' ou s'il 'refroidit'. Dans la mesure où un tel dispositif fonctionne, Frank Rosenblatt considère qu'il s'agit d'une théorie pertinente de la perception, de l'apprentissage, et de la mémoire. Conformément aux objectifs de l'intelligence artificielle, nous voici *à la fois,* en face d'une description de facultés cognitives, et d'une machine qui fonctionne. La machine propose un modèle de la *perception* en tant qu'elle s'active (elle « reconnaît ») devant certaines formes ; de l'*apprentissage* puisque grâce au processus d'essai-erreur, le dispositif devient capable d'associer les bonnes formes et son activation ; et de la *mémoire* : pour peu que le dispositif garde les valeurs prises par les potentiomètres, il sera à nouveau capable de s'activer devant les formes « apprises ». C'est une théorie différente de la « mémoire » des ordinateurs conçus selon le modèle de von Neumann, puisqu'au lieu d'être un stock d'information conservé à part, la « mémoire » est distribuée dans le réseau sous la forme de poids et de seuils d'activation. Le Perceptron propose un modèle plus « empiriste » : la mémoire n'est pas mise en scène comme un stock d'informations auquel on « pose des questions », mais comme des dispositions incorporées acquises petit à petit, par tâtonnement. C'est au fil de l'apprentissage que les neurones ajustent leurs poids, et acquièrent ainsi de la pertinence.

1.2.5. La querelle du Perceptron (1969)

La structure du Perceptron lui permet d'effectuer un certain nombre d'opérations logiques, comme le OU et le ET, suffisantes pour des problèmes de classification simple, mais pas l'opération « OU exclusif » (XOR). Pour pouvoir effectuer la fonction XOR, il faut ajouter des couches intermédiaires de neurones[130]. Mais cela rend le Perceptron trop complexe pour qu'il soit possible de trouver les bons paramètres par un tâtonnement manuel. Bien que le Perceptron obtienne de bons résultats en reconnaissance de caractères, l'impossibilité apparente de réaliser la fonction XOR jette le discrédit sur son statut de favori de la recherche en intelligence artificielle.

130 Pour une explication détaillée des problèmes liés à la fonction XOR, voir Mikel Olazaran, « A Sociological Study of the Official History of the Perceptrons Controversy » *Social Studies of Science*, vol. 26, No. 3, Août 1996, p. 625, 626.

> il s'avère que la négation de XOR – autrement dit la fonction NOT(XOR) – est pour le calcul une opération universelle, dans le sens que n'importe quelle opération binaire peut être exprimée en termes de NOT(XOR) et que, en principe, on peut assembler en entier un ordinateur à partir de puces électroniques n'effectuant que cette unique opération. Pour présenter un quelconque intérêt, un système doit être capable de détecter que deux entrées sont identiques[131].

En 1969, à l'occasion d'un livre sur le Perceptron[132], Minksy et Papert soulignent ces limites et dénoncent l'enthousiasme excessif qu'il a suscité.

> Il a été fait beaucoup de battage autour des Perceptrons, présentés comme des « machines qui apprennent » ou qui « reconnaissent les formes » et qui ont été traités comme tels dans nombre d'ouvrages, d'articles de journaux et de volumineux « rapports ». La majorité de ces écrits […] n'a aucune valeur scientifique et nous ne nous référons en général pas nommément aux travaux que nous critiquons […]. Consternés par l'influence tenace des Perceptron (et des méthodes analogues de raisonnement) sur ce qui peut être fait en matière de reconnaissance de formes, nous avons décidé d'exposer nos travaux dans un ouvrage[133].

Pour les auteurs, il s'agit de corriger cette erreur d'évaluation en clarifiant quelles sont les limites du Perceptron, sans pour autant les discréditer. Ils soulignent également que les réseaux de neurones sont un bon outil de classification, et un moyen pertinent d'étudier le fonctionnement du cerveau[134]. Mais la réception de l'ouvrage dépasse largement leurs intentions. Leurs critiques jettent un discrédit durable sur les Perceptrons et entraînent un arrêt de la plupart des recherches sur les réseaux de neurones[135]. Avec l'article de 1950 et la présentation du test, Turing inaugurait une certaine manière de penser la preuve ou l'argument dans le domaine de l'intelligence artificielle, selon laquelle il vaut mieux fabriquer des machines et tester leurs capacités que de chercher à répondre directement aux questions qui se posent. À première vue, cela permet de s'affranchir des controverses sur le sens des mots et d'apporter des « réponses » indiscutables : une machine est capable, ou non, de réaliser la tâche pour laquelle elle a été conçue. Elle passe le test, ou ne le passe pas. Ainsi, la fabrication de machines devrait constituer des arguments, voire des preuves, incontestables. Or, quelques

131 Daniel Crevier, *op. cit.*, p. 130.
132 Marvin Minsky, Seymour Papert, *Perceptrons: An Introduction to Computational Geometry*, Cambridge MA, The MIT Press, 1969.
133 *Ibid.*, p. 242, cité par Crevier, *op. cit.*, p. 131.
134 « la moitié de l'ouvrage présentait des résultats favorables au Perceptron », *Ibid.*
135 « Leur autorité était telle que l'ouvrage stoppa quasiment net la recherche en réseaux neuronaux menée au Etats-Unis. » *Ibid*.

années seulement après l'article de Turing, l'épisode du Perceptron montre, bien au contraire, que les machines sont aussi l'occasion de conflits d'interprétations. Selon les mots de Mikel Olazaran, le champ de l'intelligence artificielle n'échappe pas à la « flexibilité interprétative » (*interpretative flexibility*[136]) des résultats scientifiques. Avec une rigueur assez rare pour être soulignée, Papert et Minsky font la démarche de *reproduire l'expérience* de Rosenblatt et reprennent pas à pas les caractéristiques du Perceptron. Ils sont donc face à *la même machine.* Pourtant, leur lecture n'est pas la même. Ainsi, faute de standards sur ce qu'est une expérience *valide*[137], le type d'expérience que propose le récent champ de l'intelligence artificielle (la fabrication de machines passant des tests), échoue à mettre fin aux controverses. Le passage par la technique en multiplie au contraire les occasions : l'architecture de la machine, ses composants, la manière de programmer l'ensemble et chaque détail du programme, peuvent faire l'objet de débats. Pour le Perceptron, il aurait été possible de s'en prendre également à ses capteurs photoélectriques, de remettre en question la pertinence de la notion de « neurone », si éloignée de nos propres neurones, ou encore le choix de l'implémentation sur un Mark I, et ainsi de suite. Il faut souligner que Rosenblatt était parfaitement conscient des limitations dont souffraient le Perceptron. Il avait publié une liste de quinze problèmes non résolus et prévoyait d'essayer d'autres architectures. La controverse tenait donc moins aux limites du Perceptron qu'à la possibilité, ou à la pertinence[138], de leur trouver des solutions. À cela s'ajoute l'enjeu du financement : la popularité des réseaux de neurones menaçait de capter la majorité des investissements, au détriment des laboratoires de Papert et Minsky[139]. Aussi la controverse peut se résumer de la façon suivante : pour Rosenblatt et ses partisans, il ne s'agissait que d'un début. Les Perceptrons rencontraient des difficultés mais, avec plus de temps et plus de moyens, ils seraient en mesure de les surmonter en évoluant vers des modèles plus complexes[140]. Pour

136 « Showing the interpretative flexibility of scientific results amounts to the realization that no knowledge possesses absolute warrant, whether from logic, experiment or practice ; there can always be grounds for challenging any knowledge claim. » Olazaran, *op. cit.*, p. 611.

137 *Ibid.,* p. 612. Le problème est toujours d'actualité. François Chollet s'étonne ainsi que, plus de soixante-dix ans après la naissance du champ de l'intelligence artificielle, il n'existe aucun standard pour mesurer l'intelligence des machines. François Chollet, « On the Measure of Intelligence », arXiv:1911.01547 [cs.AI], 5 novembre 2019.

138 Pour Minsky et Papert, il ne s'agissait pas seulement d'une question de possibilité : certains des problèmes mentionnés étaient faciles à résoudre à l'aide d'algorithmes conventionnels. Olazaran, *op. cit.*, p. 631.

139 *Ibid.,* p. 628. Olazaran souligne combien Minsky et Papert avaient tort de s'inquiéter. Alors qu'ils percevaient des fonds de l'ARPA en centaines de milliers de dollars, Rosenblatt ne recevait de l'ONR que des dizaines de milliers de dollars. Il semblait bien moins « compétent » que Minsky dans la recherche d'investissements.

140 « Neural-net researchers concentrated on the positive properties of the single-layer perceptron (for example, its learning algorithm, its brain-like character, its distributed memory, its resistance to damage, its parallelism), and claimed that further research on more complex models (systems with more than one layer of adjustable connections, with connections among the units of the same layer, with backward connections, and so on) was needed in order to overcome its limitations. They were asking for time and funding to carry out that research. » *Ibid.,* p. 633.

Papert, Minsky, et leur partisans, les réseaux de neurones étaient une fausse piste. La mise au point d'algorithmes conventionnels sur ordinateur avait de meilleures chances de résoudre les mêmes problèmes. En 1950, Turing concluait un article dédié aux spéculations sur les machines intelligentes par des considérations de méthodes. De nouveau avec la controverse de 1969, le débat ne se place pas tant autour de ce qui est *vrai ou faux* (les Perceptrons sont-ils vraiment une machine qui apprend, se souvient et perçoit ?) que du chemin ou de *la méthode à suivre* (faut-il investir du temps et de l'argent dans les Perceptrons ou dans les algorithmes conventionnels ?).

Après la publication de *Perceptrons,* deux événements viennent enfoncer les clous dans le cercueil des réseaux de neurones : Frank Rosenblatt meurt et l'ARPA choisit de soutenir officiellement l'école symbolique (les partisans de Minsky et Papert)[141], ce qui lui donne un accès privilégié aux investissements et aux installations informatiques – rares à l'époque. Aussi la controverse est-elle considérée comme close et « l'histoire officielle » retient que Minsky et Papert ont défait Rosenblatt en démontrant que les réseaux de neurones étaient une fausse piste. Mais la « démonstration » de Minsky et Papert est tout aussi contestable que l'étaient, selon eux, les travaux de Rosenblatt. Comme le souligne Olazaran, la « flexibilité interprétative » laissait une place pour que les choses se passent autrement, et de fait, elles finissent par prendre un nouveau cours. Contrairement à « l'histoire officielle », les travaux sur les réseaux de neurones se poursuivent, via notamment des chercheurs se réclamant de champs voisins mais distincts de l'intelligence artificielle (sciences cognitives, psychologie, neurosciences, physique). Les progrès effectués dans les années 1980 redonnent du crédit aux réseaux de neurones et amènent la communauté de l'intelligence artificielle à réinterpréter la lecture des Perceptrons faite par Minsky et Papert. En 1988, Minsky lui-même semble admettre qu'il a eu tort[142]. Ainsi, la controverse du Perceptron, et son analyse par Mikel Olazaran, révèle une certaine *contingence* du sens donné aux machines fabriquées par les chercheurs en intelligence artificielle. En fonction des succès ou des déconvenues académiques des uns et des autres, une « histoire officielle » s'écrit, qu'il faut réécrire quelques années plus tard lorsque de nouveaux résultats, et la formation autour d'eux d'une communauté de chercheurs capable d'attirer des financements, viennent changer la donne.

Dès l'article fondateur de Turing, le projet d'intelligence artificielle véhicule l'ambition de s'affranchir de la contingence : remplacer le débat par la fabrication de machines permettrait d'apporter des réponses définitives et de mettre tout le monde d'accord. Mais dès la querelle du

141 Principalement à travers l'Information Processing Techniques Office (IPTO), dirigé par Licklider.
142 Olazaran, *op. cit.,* p. 652 note 26.

Perceptron, il apparaît que, loin d'y mettre fin, le choix de passer par les machines n'a fait que multiplier les sujets de controverse.

1.2.6. L'école symbolique

Dans les années 1970, avant que l'histoire ne montre que les propos de Minksy et Papert pouvaient être remis en question, « l'histoire officielle » de l'intelligence artificielle présente leurs résultats comme définitifs : ils ont « démontré » qu'il est impossible d'entraîner des réseaux de neurones à plusieurs couches, et donc que les réseaux de neurones ne permettront pas de fabriquer des machines intelligentes. Selon Mikel Olazaran, si « l'histoire officielle » présente les travaux de Minsky et Papert comme indiscutables, c'est parce que se produit un mouvement de crédit, dans les deux sens du terme, en direction de l'école symbolique[143]. L'institutionnalisation de l'école symbolique est un tel succès qu'elle en vient à apparaître comme *la seule méthode valide* en intelligence artificielle. Newell et Simon en tirent même un argument en faveur de l'école symbolique : cela doit être la bonne méthode, puisqu'aucune approche alternative ne réussit à lui opposer de concurrence[144]. Pourtant, l'approche symbolique se définit principalement par opposition à la direction que prenaient les réseaux de neurones. Alors que ces derniers prenaient comme point de départ l'imitation de processus biologiques inconscients, l'école symbolique préfère identifier et répliquer les processus mentaux conscients. Il s'agit de déterminer comment s'enchaînent les idées et de traduire cet enchaînement sous la forme de programmes. Cela revient également à privilégier la programmation des ordinateurs, plutôt que la fabrication de machines *ad hoc* : la machine peut être programmée, ou « amenée à » (« made to ») avoir un comportement intelligent, plutôt que fabriquée pour l'occasion (« made to »), à l'instar du Mark 1 de Rosenblatt[145]. Avec ce parti

143 « According to the official history, Minsky and Papert replied to Rosenblatt's overclaiming and showed that progress in neural nets was not possible - and after that this field was largely abandoned. But if, as I have shown here, Minsky and Papert did not quite show that, and if (as I will point out soon) neural nets were not completely abandoned, what was the role of the official history? It is my view that its role can only have been the legitimation of the emergence and institutionalization of the symbolic approach, which came to be seen as the 'right' approach to Al, and as occupying the whole Al discipline. » Olazaran, *op. cit.*, p. 640.

144 « In the 1970s, symbolic Al's leading researchers used the 'we are the only Al paradigm' argument in their rhetoric, as can be seen in this quote from a seminal paper by Newell and Simon: « The principal body of evidence for the symbolic hypothesis that we have not considered [so far in this paper] is negative evidence: the absence of specific competing hypotheses as to how intelligent activity might be accomplished whether by man or by machine. » Olazaran, *op. cit.*, p. 640.

145 La fabrication de machines spécialisées continue cependant, notamment celle de robots, et en particulier avec l'approche « micro-worlds ».

pris, l'école symbolique bénéficie de l'engouement pour les ordinateurs. À défaut de les rendre intelligents, leurs travaux permettent au moins de faire avancer la recherche et l'innovation à leur sujet – légitimant d'autant plus les financements reçus. C'est le cas du langage de programmation LISP, qui apparaît à la fois comme une direction de recherche pour la communauté des chercheurs en intelligence artificielle, et comme un outil informatique au service de la programmation en général. L'école est dite « symbolique » car elle met l'accent sur l'utilisation par l'ordinateur de « symboles » (au sens où ils « représentent » un élément du monde, une opération, une partie de la machine…) et sur la mise au point de règles de manipulation de ces symboles[146]. Les éléments suivants sont généralement mis en avant pour distinguer l'école symbolique de son école rivale, baptisée « connexionniste[147] » en référence à l'accent mis sur le rôle des connexions entre neurones :

- l'**apprentissage**, ou plutôt l'ajout d'une connaissance, se fait directement dans le cas d'un système symbolique : il suffit d'écrire une nouvelle règle ou de définir un objet, qu'on rajoute au programme. Dans le cas d'un réseau de neurones, elle se fait indirectement, par la présentation d'exemples sur lesquels la machine se calibre. Elle ajuste ses paramètres de manière à ce que ceux-ci encodent les points communs à la série d'exemples (et de contre-exemples), points communs constituant autant de caractéristiques définissant l'objet à reconnaître par le réseau. Autrement dit, dans l'école symbolique, le programmeur ajoute directement des règles explicites. Tandis que dans l'école connexionniste, il montre des exemples qui servent à paramétrer le réseau. Pour l'école symbolique, le savoir doit être préalablement enregistré puis manipulé par un ensemble de règles explicites, tandis que pour l'école connexionniste, il doit être acquis par interaction avec l'environnement. Cette opposition a pu être comparée à l'opposition philosophique entre rationalistes et empiristes, et à sa traduction dans le débat entre inné et acquis : l'école symbolique repose sur des savoir « innés », ou ajoutés par le programmeur, tandis que l'école connexionniste met l'accent sur leur acquisition via la perception ;
- la **représentation**. Cela se traduit en une théorie différente de la représentation mentale. Dans un système symbolique, on dispose de définitions explicites, tandis que dans un réseau de

146 Voici la définition qu'en donne Olazaran : « Within symbolic AI, intelligence and cognition are seen as processes of symbol manipulation and transformation. A symbolic system relies on its representational structures and on the possibility of applying structure-sensitive operations to them. Representational structures are manipulated and transformed according to certain rules and strategies (embodied in computer programs), and the resulting expression is the solution to a given problem. » *op. cit.*, p. 614.

147 C'est Hebb qui utilise le terme pour la première fois en 1949, repris ensuite par Rosenblatt en 1958. Daniel Andler, « From paleo to neo-connectionism », in G. Van der Vijver (ed.), *Perspectives on Cybernetics*, Dordrecht, Kluwer, p. 125-146.

neurones, la représentation de l'objet interne tient à l'ensemble des poids et seuils lui correspondant : c'est un ensemble de valeurs numériques ;

- la **mémoire.** Dans un système symbolique, la mémoire consiste en un ensemble d'informations stockées à part, que le système vient interroger lorsqu'il en a besoin. Dans un réseau de neurones, la mémoire est distribuée en tous points du réseau : chaque poids ou seuil pris aux différents endroits du réseau encode une partie de la représentation acquise ;

- **les opérations.** Dans un système symbolique, les opérations sont **séquentielles,** les symboles passent par une série de manipulations successives. Tandis que dans un système connexionniste, les opérations sont en partie **parallèles** : à chaque couche, les neurones opèrent simultanément leurs calculs respectifs.

École	**Symbolique**	**Connexionniste**
Apprentissage	Ajout par le programmeur	Par l'exemple
Représentation	Définitions explicites	Valeurs numériques
Mémoire	Stockée à part	Distribuée dans le réseau
Opérations	Séquentielles	Parallèles

Les deux écoles ont pour horizon un modèle informatique représentant et reproduisant l'intelligence, mais l'école symbolique part d'une certaine idée du *mental,* tandis que l'école connexionniste part d'une certaine idée du *corps* pensant : le cerveau comme machine se calibrant via les perceptions[148]. L'école symbolique met l'accent sur la manipulation d'idées, tandis que l'école connexionniste s'appuie sur la base matérielle de ces idées, sur ce à quoi elles correspondent dans le cerveau. Pour l'école dite symbolique, l'intelligence artificielle doit consister en des programmes effectuant de la logique. Pour l'école connexionniste, il faut fabriquer des machines apprenantes, éventuellement en imitant la structure du cerveau.

Si la distinction symbolique / connexionniste semble opératoire, elle est, là encore, sujette à une certaine « flexibilité interprétative ». Avant de revenir sur les raisons qui permettent de ne pas la prendre pour argent comptant, soulignons *l'intérêt* qu'il pouvait y avoir, du point de vue de l'école symbolique, à ce que cette distinction soit bien marquée, pour bien flécher les financements.

148 Pour les besoins de l'exposé, nous simplifions la distinction entre école symbolique et école connexionniste. Le sujet est délicat, tant à cause de la variété des approches *au sein* des écoles symboliques et connexionnistes, de leurs évolutions respectives au cours de l'histoire de l'intelligence artificielle, que de la « flexibilité interprétative » de la distinction. Pour certains auteurs, elle n'a pas lieu d'être, tandis que pour d'autres elle constitue une fracture décisive. Voir Daniel Andler, « Connexionnisme et cognition : A la recherche des bonnes questions », *Revue de synthèse,* Volume 111, 1-2, janvier 1990, p. 95–127.

Cinquante ans plus tard, l'intérêt est toujours aussi fort, mais il s'est déplacé. Dans les années 1970, l'école symbolique avait intérêt à ce que la distinction soit marquée pour exclure les laboratoires rivaux et capter les financements. Dans les années 2010-2020, la situation est la même, mais elle s'est inversée, les financements allant principalement aux projets affiliés à l'apprentissage profond (descendants des réseaux de neurones). En quarante ans, les rivaux exclus par l'écriture de « l'histoire officielle » se sont rassemblés en une communauté alternative et ont inversé le rapport de force, reprenant à leur compte la distinction symbolique / connexionniste, qui s'était constituée à leur détriment, pour exclure à leur tour leurs rivaux.

1.2.7. Le retour des réseaux de neurones (1979-1986)

Contrairement aux affirmations exagérées de la « science officielle », les recherches sur les réseaux de neurones n'ont pas été stoppées par la controverse du Perceptron. Dans les années 1970, les laboratoires d'intelligence artificielle s'orientent majoritairement vers une approche symbolique, mais les réseaux de neurones continuent à faire l'objet de travaux, notamment en neurosciences, en psychologie, ou en sciences cognitives.

Nous avons vu comment, en liant son destin à celui des ordinateurs standards, l'école symbolique avait renforcé son attractivité pour les financeurs. À partir de la fin des années 1970 et dans les années 1980, ce lien joue en sa défaveur. L'accès aux ordinateurs est de plus en plus facile, et de nouveaux arrivants peuvent se lancer dans des recherches en intelligence artificielle sans avoir à s'affilier aux laboratoires dominants.

Dans les années 1980, des chercheurs venus d'horizons différents forment une communauté concurrente autour des réseaux de neurones[149]. En 1979, Geoffrey Hinton et James Anderson organisent une conférence sur le sujet à La Jolla, en Californie, qui aboutit au livre *Parallel Models of Associative Memory* (1981). Dans la foulée, les psychologues David Rumelhart et James McClelland, forment un groupe de recherche à l'Université de San Diego : le *Parallel Distributed Processing* group (ou *PDP*). Pour les membres de ce groupe, si l'activité de nos neurones arrive à des résultats si complexes, c'est qu'ils doivent fonctionner de manière parallèle – étant donné la lenteur de chaque neurone. Les neurones se répartissent les calculs et effectuent ceux-ci *simultanément*, ce qui les distingue des ordinateurs de von Neumann, où la

149 Nous résumons le récit qu'en donne Olazaran, *op. cit.*, p. 643-645.

séparation entre la mémoire et le processeur impose une séquentialité des opérations[150]. Le travail et les publications du groupe PDP suscitent de nouvelles vocations et attirent des chercheurs chevronnés provenant d'autres disciplines. S'inspirant du comportement des matériaux magnétiques, le physicien John Hopfield propose un modèle de réseau de neurones qui apporte une théorie plausible de mémoire associative[151]. Dans le prolongement, Hinton et Sejnowski développent les « machines de Boltzmann », un réseau de neurones dérivé des réseaux de Hopfield, pour lequel ils mettent au point un algorithme d'apprentissage[152]. Ils démentissent ainsi l'argument utilisé contre le Perceptron, et plus généralement contre les réseaux de neurones, selon lequel il est impossible d'entraîner un réseau multicouche. Les membres du groupe PDP ne sont pas les seuls à travailler dans ces directions. Dès 1980, Kunihiko Fukishima présentait, avec le *neocognitron*, un réseau multicouches inspiré du système nerveux de la vision (plus précisément des travaux de Hubel et Wiesel), pour lequel il propose plusieurs algorithmes d'entraînement[153]. En 1986, Rumelhart, Hinton et Williams présentent l'algorithme de *back-propagation*[154], permettant d'entraîner un réseau de neurones multi-couches de même type que le Perceptron[155]. Ils marquent ainsi le retour effectif des réseaux de neurones dans le champ de l'intelligence artificielle et rouvrent le débat entre école symbolique et école connexionniste.

1.2.8. L'algorithme de *back-propagation* (1986)

Avec l'invention de l'algorithme de *back-propagation*[156], Rumelhart, Hinton et Williams mettent au point un moyen d'entraîner des réseaux de neurones à plusieurs couches, surmontant

150 Cela n'empêche pas de *simuler* une architecture parallèle sur des ordinateurs séquentiels. La recherche sur les réseaux de neurones n'impose pas de se doter du *hardware* correspondant.

151 John J. Hopfield, « Neural Networks and Physical Systems with Emergent Collective Computational Abilities », *Proceedings of the National Academy of Sciences*, vol. 79, p. 2554-2558, avril 1982.

152 Ackley, Hinton, et Sejnowski, « A learning algorithm for Boltzmann machines », *Cognitive Science*, 9 (1), 1985, p. 147–169.

153 Kunihiko Fukushima, « Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position », *Biological Cybernetics* 36, 193–202, 1980.

154 Rumelhart, Hinton et Williams, « Learning representations by back-propagating errors », *Nature*, n° 323, 1986, p. 533-536.

155 C'est-à-dire avec une entrée et une série d'étapes (couches de neurones) amenant à une sortie (« feed forward »), alors que dans une machine de Boltzmann, tout comme les réseaux de Hopfield, les neurones sont tous connectés les uns aux autres.

156 L'invention est conventionnellement attribuée à Rumelhart, Hinton et Williams, « Learning representations by back-propagating errors », *op. cit.* Dans l'article, les auteurs reconnaissent que d'autres chercheurs sont arrivés

les difficultés identifiées par Minsky et Papert en 1969. Pour en illustrer le fonctionnement, voici une description simplifiée de l'apprentissage d'une tâche de classification d'images[157]. Chaque pixel des images est représenté par trois nombres correspondant à la valeur des couleurs fondamentales (rouge, bleu, vert). Le but de l'algorithme est d'associer cette suite de nombres à un mot, celui qui décrit correctement ce que contient l'image (par exemple « un chat »), choisit parmi une liste d'objets, les « classes » : chat, chien, poisson…

Pour peu que les classes de la base de données soient consistantes (qu'il existe une structure commune à tous les objets de la même classe), on suppose qu'il existe une fonction qui associe les ensembles de pixels aux bonnes classes. Le but de l'algorithme de *back-propagation* est d'amener le réseau de neurones à *approximer cette fonction* de façon à ce que, si on lui donne une nouvelle image, il sache lui attribuer la bonne classe.

Entre le nombre d'entrée et le mot de sortie, il y a un certain nombre de neurones organisés en couches. Comme pour les neurones du Perceptron de Rosenblatt, chaque neurone affecte un **poids** à chaque valeur reçue qui encode l'importance donnée à cette information. Il additionne l'ensemble des valeurs multipliées par leurs poids et ajoute au résultat un dernier nombre, appelé le **biais.** Suivant la valeur prise par l'ensemble (le **seuil**), il envoie ou non cette information à la couche suivante.

Autrement dit chaque neurone est une fonction mathématique qui reçoit plusieurs *inputs* (les valeurs d'un certain nombre de pixels, ou les valeurs transmises par d'autres neurones), en effectue la somme pondérée et, si un certain seuil est atteint, produit un *output*. Cet output constitue l'*input* de la couche de neurones suivante.

à des résultats similaires de manière indépendante : « Variants on the learning procedure have been discovered independently by David Parker (personal communication) and by Yann LeCun. »

157 Nous emprutons les étapes de cette description à la vulgarisation qu'en fait Geoffrey Hinton : « The Foundations of Deep Learning », *Youtube,* 7 février 2018.

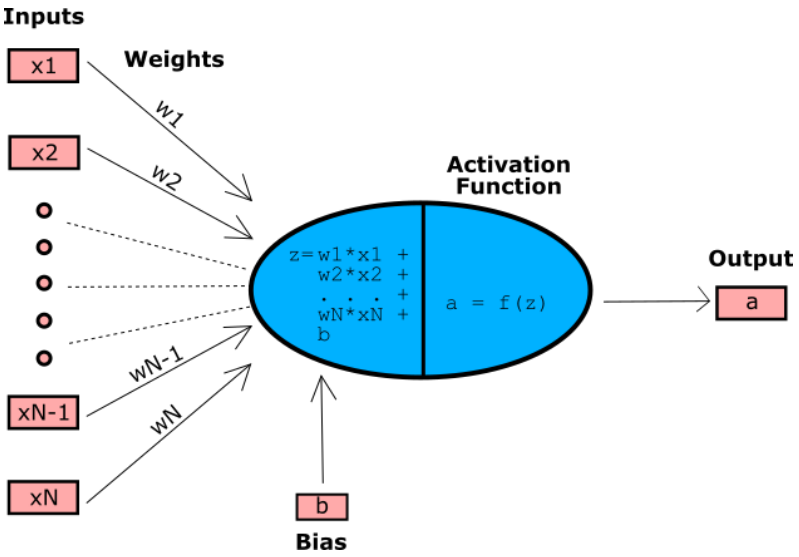


Schéma d'un neurone recevant N inputs xN auxquels sont ajoutés les poids (w pour weights) ainsi qu'un biais. Le neurone effectue la somme pondérée des inputs et y ajoute le biais pour calculer l'output, qui servira d'input aux neurones suivants. Image : James Fulton, Towards data science, https://towardsdatascience.com/emulating-logical-gates-with-a-neural-network-75c229ec4cc9 page consultée le 20 septembre 2020.

Tout comme le Perceptron de Rosenblatt, les poids et les biais du réseau de neurones sont d'abord attribués au hasard, le réseau part d'une combinaison de paramètres aléatoires. L'enjeu est d'adapter les paramètres (poids et biais) de façon à ce que soient associées les bonnes suites de nombres (images) avec les bons mots (classes) : il s'agit de **trouver la meilleure combinaison de paramètres**[158]**.** On soumet au réseau de neurones une **base de données d'entraînement,** c'est-à-dire un grand nombre d'exemples qui ont été « corrigés » au préalable : la bonne réponse est connue et indiquée dans un label qui accompagne l'image. Le réseau de neurones calcule son **taux d'erreur,** l'écart entre ses réponses et le label de chaque image.

Le but de l'algorithme de *back-propagation* est de **traduire cet écart à la bonne réponse en une modification des paramètres** qui diminue cet écart. Pour chaque poids, il teste l'effet d'une légère modification et conserve les modifications qui améliorent son résultat. Il y a une « rétro-propagation » du taux d'erreur en une adaptation des paramètres – « rétro-propagation » qui donne son nom à l'algorithme.

158 « The aim is to find a set of weights that ensure that for each input vector the output vector produced by the network is the same as (or sufficiently close to) the desired output vector. » Rumelhart, Hinton et Williams, *op. cit.*

1.2.9. Machines inductives et représentations mentales

Lorsqu'il trouve une combinaison de paramètres qui minimise le taux d'erreur, on dit que l'algorithme **converge**. Si le réseau s'est correctement paramétré, certaines informations des images déclenchent l'attribution de la bonne classe (« s'il y a tel type de moustaches, c'est probablement un chat »).

> Au contact d'un grand nombre d'exemples de configurations complètes […], le système s'adapte aux régularités de l'environnement en ajustant ses poids synaptiques, ce qui lui permet d'une part de réagir sans aucune erreur aux exemples qui lui ont été présentés au cours de l'apprentissage, d'autre part de réagir « intelligemment » à d'autres configurations incomplètes – soit en les assimilant à des parties de configurations connues, soit en y discernant un mélange de configurations connues, et en les complétant en conséquence. Bref, le système se comporte en détecteur de régularités statistiques multidimensionnelles[159].

Le réseau de neurones fonctionne comme s'il **extrayait des caractéristiques** de l'image de façon à les associer à une classe[160]. La première couche identifie certaines caractéristiques fondamentales (des angles, des surfaces unies, des contours…), à partir desquelles les couches suivantes pourront identifier des caractéristiques plus élaborées (un œil, une moustache), qui seront communiquées à la dernière couche pour que celle-ci attribue une probabilité à chaque classe (s'il y a une moustache et tel type d'œil alors voici la probabilité que cela soit un chat[161]). Ainsi, il n'est pas nécessaire d'indiquer au réseau de neurones quelles sont les caractéristiques qui permettent de reconnaître les objets[162]. Les réseaux de neurones ainsi utilisés sont donc des

159 Daniel Andler, *op. cit.*, p. 103.

160 « As a result of the weight adjustments, internal 'hidden' units which are not part of the the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. » Rumelhart, Hinton et Williams, *op. cit.*

161 Cette organisation hiérarchique permet un « recyclage » des réseaux de neurone appelé le *transfer learning*. Il est possible d'enlever la dernière couche et de réutiliser un réseau pour reconnaître une autre série d'objets dans la mesure où les caractéristiques de base sont les mêmes (lignes, contours…).

162 Avant que les réseaux de neurones ne s'imposent dans le champs de la vision artificielle, il fallait définir en amont les caractéristiques à identifier dans l'image. Hubert Dreyfus commente ainsi les premiers programmes de reconnaissance de formes : « Tous ces programmes comportent une recherche des traits caractéristiques des éléments présentés, et leur comparaison avec les 'définitions', préalablement apprises ou incorporées, correspondant à chacun des caractères à identifier. Toute l'astuce est de parvenir à trouver les traits caractéristiques qui permettront l'identification à coup sûr, autrement dit, des traits qui se retrouvent dans tous les cas, indépendamment des variations de taille, d'orientation, ou d'autres distorsions. » Hubert Dreyfus, *Intelligence artificielle : mythes et limites*, Paris, Flammarion, 1984. (traduit de l'anglais par Rose-Marie Vassallo-Villaneau avec le concours de Daniel Andler, *What Computers Can't Do, The Limits of Artificial Intelligence* New-York, Harpers & Row, 1979 (1è éd. 1972), p. 45.

machines inductives. Elles **décèlent une structure dans le désordre des données qui lui sont fournies**. Elles « apprennent » à « identifier et à démêler les facteurs explicatifs sous-jacents cachés dans le milieu observé des données sensorielles de bas niveau[163] ». Ces structures sont encodées par la combinaison de paramètres du réseau de neurones, que l'on peut représenter sous la forme d'un vecteur. Pour Rumelhart, Hinton et Williams, le réseau de neurones « se construit une représentation interne appropriée[164] ». C'est pour eux une différence majeure avec les Perceptrons. Ceux-ci étant réglés à la main, il n'y a pas « d'apprentissage des représentations[165] ». L'algorithme de *back-propagation* fournit un modèle qui permet de prendre une position ferme dans les controverses qui agitent les sciences cognitives. Dans la cognition selon les réseaux de neurones, les concepts se forment à partir des données de la perception et donnent lieu à des « représentations internes », des états mentaux équivalents à l'état des neurones et de leurs connexions – état qui peut être représenté par un vecteur. Plus tard, Hinton pourra ainsi affirmer avec aplomb qu'« une pensée, c'est juste un sacré gros vecteur d'activité neurale[166]. » Le cerveau serait une machine à fabriquer des représentations, et il y aurait *équivalence entre une pensée et l'état du cerveau à un moment donné* (la combinaison des paramètres du réseau de neurones).

La conclusion de l'article de 1986 en reste toutefois à une interprétation prudente : l'algorithme de *back-propagation* et la recherche d'une combinaison de paramètres minimisant le taux d'erreur ne correspondent probablement pas au fonctionnement effectif du cerveau, mais n'en constituent pas moins un modèle élégant de construction de « représentations internes » et une bonne direction de recherche[167].

---

163 Yosha Bengio, A. Courville, P. Vincent, « Representation Learning: A Review and New Perspectives », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, n°8 et al. Cité par Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones, L'invention des machines inductives et la controverse de l'intelligence artificielle », in *Réseaux*, n°211, 2018 / 5, p. 173-220.

164 « The learning procedure must decide under what circumstances the hidden units should be active in order to help achieve the desired input-output behavior. This amounts to deciding what these units should represent. We demonstrate that a general purpose and relatively simple procedure is powerful enough to construct appropriate internal representations. » Rumelhart, Hinton et Williams, *op. cit.*

165 « In perceptrons there are feature analysers between the input and output that are not true hidden units, because their input connections are fixed by hand, so their states are completely determined by the input vector : they do not learn representations. » Rumelhart, Hinton et Williams, *op. cit.*

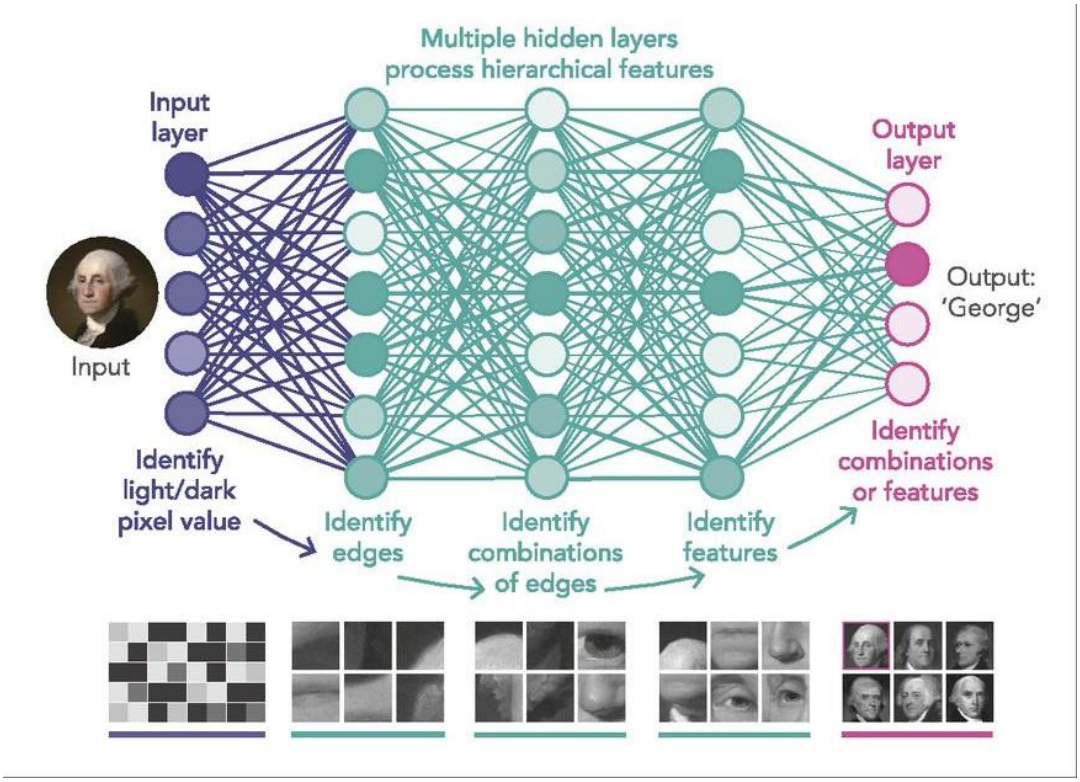
166 « what a thought is, is just a great big vector of neural activity. ». Geoffrey Hinton, interviewé par Andrew Ng, « Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton », https://www.youtube.com/watch?v=-eyhCTvrEtE, page consultée le 4 octobre 2019.

167 « The learning procedure, in its current form, is not a plausible model of learning in brains. However, applying the procedure to various tasks shows that interesting internal representations can be constructed by gradient descent in weight-space, and this suggests that it is worth looking for more biologically plausible ways of doing gradient descent in neural networks. » Rumelhart, Hinton et Williams, *op. cit.*

1.2.10. Réseaux convolutionnels et cortex visuel

En s'inspirant du fonctionnement du cortex visuel du chat tel qu'il a été décrit par Hubel et Wiesel[168], Yann LeCun et son équipe mettent au point un type particulier de réseau de neurones appliqués à la reconnaissance de forme, les réseaux convolutionnels.

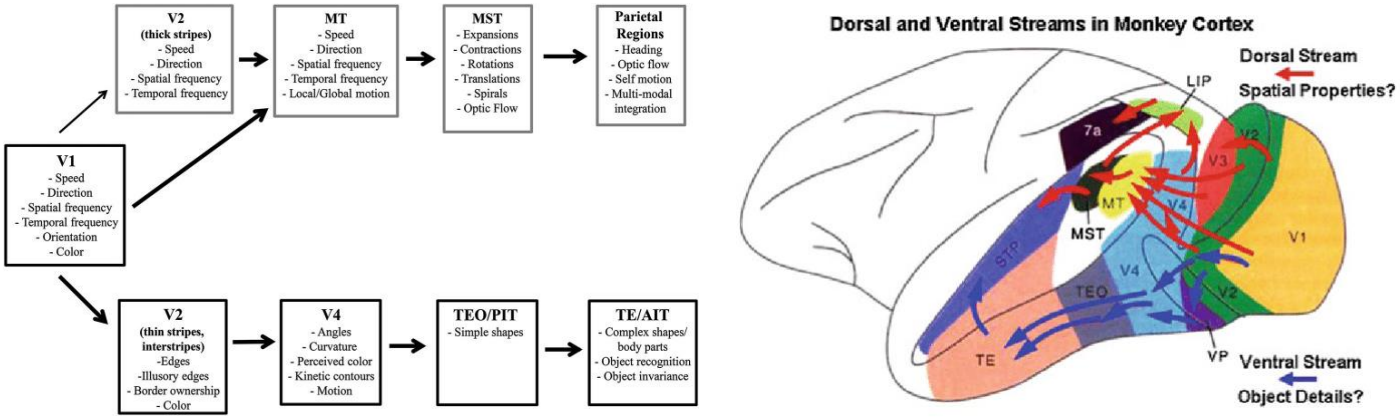
Nous avons déjà évoqué l'aspect **hiérarchique** des réseaux de neurones : les premières couches détectent certaines caractéristiques élémentaires (contraste, dégradé) et envoient les informations à la couche suivante où les neurones compilent ce qui a été détecté pour en déduire la présence de caractéristiques plus élaborées (lignes, frontières…). Le même processus est répété de couche en couche, permettant de détecter des objets à la complexité croissante.



À partir des valeurs des pixels (couche d'input), le réseau identifie des caractéristiques de plus en complexes à partir de la combinaison des caractéristiques élémentaires repérées. Source Waldrop, M Mitchell, « News Feature: What are the limits of deep learning? » Proceedings of the National Academy of Sciences of the United States of America, *vol. 116, 4, 2019, pages 1074-1077. doi:10.1073/pnas.1821594116*

168 David Hubel et Torsten Wiesel. « Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. » *The Journal of Physiology* 160, 1962.

Cet aspect hiérarchique est un point commun important entre les réseaux de neurones et le système visuel des mammifères. Dans ce dernier, les premières couches de neurones réagissent à des caractéristiques très précises comme les ombres et contrastes (zone V1), et se combinent dans des formes de plus en plus élaborées : lignes et frontières en V5, angles et couleurs dans le cortex inférotemporal postérieur[169].



À gauche : à partir du signal visuel, les aires successives du cerveau construisent une représentation de plus en plus élaborée. À droite : carte simplifiée d'un cerveau de singe représentant les aires mentionnées. Sources : Perry, Carolyn & Fallah, Mazyar, « Feature Integration and Object Representations along the Dorsal Stream Visual Hierarchy » Frontiers in computational neuroscience, *8, 2014, 84. Aine, Cheryl & Supek, Selma & Sanfratello, Lori & Stephen, Julia, « Selection of Stimulus Parameters for Visual MEG Studies of Sensation and Cognition »*. Magnetoencephalography: From Signals to Dynamic Cortical Networks, *2012, p. 767-799.*

Un autre élément clef tiré des travaux de Hubel et Wiesel est la **distinction entre cellules simples et cellules complexes**. À chaque couche, des « cellules simples » détectent les caractéristiques pour une position et une orientation particulière, tandis que des « cellules complexes » retiennent l'information en faisant abstraction de la position et de l'orientation. Intercaler des cellules simples et des cellules complexes permet de détecter des caractéristiques de plus en plus élaborées, indépendamment des variations contextuelles[170].

169 « Early DCNN were inspired by Hubel & Wiese's discovery that early ventral stream neurons seemed to be sensitive to very specific and local features like shadings and contrasts in V1, which was later enhance with a variety of imaging methods to suggest a whole processing cascade, from lines and borders in V5, to angles and colors in TEO/PIT (posterior inferotemporal), to figures and objects in TE/AIT (anterior inferotemporal). » Cameron Buckner, « Deep Learning: A Philosophical Introduction », *Philosophy Compass,* 14, 2019.

170 Yann Lecun, Koray Kavukcuoglu, Clement Farabet, « Convolutional Networks and Applications in Vision », *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*. 253-256, 2010. C'est Kunihiko Fukushima qui s'inspire le premier des découvertes de Hubel et

Ainsi, les réseaux convolutionnels s'ajustent systématiquement aux multiples variations qui entravaient la reconnaissance d'images (variation de taille, de pose, de localisation, d'orientation…) ou de sons (variations d'intensité, de ton, de prononciation, de durée…). Cela explique également leur succès dans d'autres tâches où il faut évaluer une situation « en gros » plutôt que dans le détail, comme le jeu de Go[171]. Autrement dit, les réseaux convolutionnels arrivent à surmonter la très grande difficulté que posent les innombrables situations où il faut reconnaître une forme alors qu'elle n'est là qu'« à peu près ».

1.2.11. L'engouement connexionniste

L'année 1986 marque le retour des réseaux de neurones. Le groupe PDP publie une somme en deux volumes, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*[172], qui expose la nouvelle assise théorique du mouvement connexionniste et lui assure une certaine notoriété. Ce qui est maintenant désigné comme l'« école connexionniste » dispose désormais de suffisamment d'arguments pour se mesurer à l'école symbolique. Grâce aux efforts de Rumelhart et McClelland, la DARPA se laisse convaincre d'investir dans la recherche sur les réseaux de neurones, entraînant à sa suite les agences européennes et japonaises[173].

À la fin des années 1980, une série d'inventions permettent de montrer que la théorie connexionniste donne lieu à des algorithmes qui marchent. Rumelhart et McClelland ont mis au point un réseau de neurones capable de former le prétérit d'un verbe anglais à partir de l'infinitif[174]. En 1990, lorsqu'il présente le connexionnisme au public francophone, Daniel Andler fait part de son étonnement :

Wiesel : « Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position », *Biological Cybernetics* 36, 1980, p. 193–202.

171 « Go strategy, for example, should also be tolerant to small changes in the position or rotation of stone placements patterns […] ». Cameron Buckner, « Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks », *Synthese*, 12, p. 1-34, 2018.

172 Rumelhart, David, David McClelland, PDP Research Group, *Parallel Distributed Processing, Volume 1. Explorations in the Microstructure of Cognition : Foundations*, Cambridge MA, MIT Press, 1986. Et *Parallel Distributed Processing, Volume 2. Explorations in the Microstructure of Cognition : Psychological and Biological Models*, Cambridge MA, MIT Press, 1986.

173 Olazaran, *op. cit.*, p. 646.

174 Parallel Distributed Processing, n. 1, vol. 2, chap. 18, p. 216-271 : « On learning the past tenses of English verbs ». L'exemple est mentionné par Daniel Andler, « Connexionnisme et cognition: A la recherche des bonnes questions », *Revue de synthèse*, Volume 111, 1-2, janvier 1990, p. 102-103.

> L'algorithme est indépendant de la fonction que le système doit apprendre, et son application n'exige pas l'intervention du modélisateur (c'est vraiment un algorithme !). Quant au corpus, il est vaste, mais non exhaustif : une proportion non négligeable de verbes, tant réguliers qu'irréguliers, n'y figure pas. Le système est capable de maitriser le corpus, au terme d'une (longue) période d'apprentissage ; après quoi il conjugue aussi presque infailliblement tout autre verbe anglais. Il est essentiel de remarquer qu'aucune règle n'est enseignée ou indirectement fournie au système par le modélisateur (en revanche, celui-ci est entièrement responsable du « pré-traitement » conduisant la représentation phonémique, ainsi que du choix du corpus et du protocole d'apprentissage)[175].

D'autres algorithmes impressionnent le public savant et moins savant : Sejnowski et Rosenberg fabriquent NETtalk, un réseau de neurones capable de « lire » l'anglais « à haute voix », c'est-à-dire de produire le son correspondant à un texte (« text-to-speech[176] »). Jeffrey Elman présente un réseau capable d'« encoder la structure du langage » et de compléter les phrases qu'on lui soumet[177]. La reconnaissance d'images n'est pas en reste : Yann LeCun et son équipe mettent au point un système de reconnaissance des chiffres (LeNet)[178], suffisamment fiable, fait remarquable[179], pour être commercialisé aux États-Unis et en France : « À la fin des années 1990, [ce] système lit entre 10 et 20 % de tous les chèques émis aux États-Unis. C'est l'un des plus spectaculaires succès des réseaux de neurones de cette décennie[180]. »

Ainsi, dès la fin des années 1980, les réseaux de neurones font leurs preuves dans la manipulation de texte, de son et d'image. En 1992, alors qu'il publie la troisième édition de son ouvrage si critique envers le projet d'intelligence artificielle, Hubert Dreyfus y ajoute une longue introduction où il se demande si ses attaques ne seront pas rendues obsolètes par

175 Daniel Andler, *op. cit.*, p. 103.

176 Terrence Sejnowski et Charles Rosenberg, « NET talk: A parallel network that learns to read aloud », *The Johns Hopkins University EE and CS Technical Report*, Janvier 1986.

177 Jeffrey Elman, « Distributed Representations, Simple Recurrent Networks, And Grammatical Structure », *Machine Learning*, 7, p. 195–225, 1991. Nous revenons sur la notion d'encodage de la structure du langage lors du commentaire de l'article de Juan Luis Gastaldi à ce sujet.

178 LeCun, Boser, Denker, Henderson, Howard, Hubbard, Jackel, « Backpropagation Applied to Handwritten Zip Code Recognition », *Neural Computation*, vol. 1, n° 4, 1989, p. 541-551.

179 On ne souligne pas assez la distance qu'il y a entre une expérience de laboratoire et un système fiable utilisable par l'industrie. Entre les deux, la NASA distingue ainsi neuf niveaux dits de « technology readiness level », chacun pouvant demander un travail considérable : il faut passer de l'expérience au système et rendre le système robuste aux aléas de l'environnement et de l'usage. Jim Banke, « Technology Readiness Levels Demystified », 20 août 2010, https://www.nasa.gov/topics/aeronautics/features/trl_demystified.html, page consultée le 12 mars 2021.

180 Yann LeCun, *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond,* Paris, Odile Jacob, 2019, p. 58-59.

l'approche connexionniste – peut-être ne s'appliquent-elle qu'à l'école symbolique, maintenant qualifiée, avec affection ou dédain, la « Good Old Fashioned AI » (GOFAI)[181].

1.2.12. Les ripostes de l'école symbolique

Les succès de l'approche connexionniste étant soumis, comme le Perceptron de Rosenblatt, à la « flexibilité interprétative », les tenants de l'école symbolique ne se privent pas de la critiquer. Steven Pinker et Alan Prince reprochent aux systèmes connexionnistes leur dépendance au corpus d'exemples et au protocole d'apprentissage[182]. Fodor et Pylyshyn, dans un article qui fait date[183], « estiment que le connexionnisme est fondamentalement inadéquat comme théorie de la cognition », puisqu'« il ne fait pas place aux représentations structurées, lesquelles sont seules susceptibles d'expliquer un aspect central de la cognition[184]. » Autrement dit, ou bien le connexionnisme arrive à modéliser une « couche » de raisonnement, et il n'est dans ce cas qu'une variante de l'école symbolique, une manière d'approcher la cognition « par le bas » (l'architecture du cerveau) plutôt que « par le haut » (les opérations mentales) ; ou bien le connexionnisme en reste à l'architecture matérielle et dans ce cas-là il n'« explique » rien – ce n'est pas de l'intelligence artificielle.

À ces réticences théoriques s'ajoutent des problèmes plus pratiques, au premier rang desquels figure le « problème de la convexité », un défaut de l'algorithme de *back-propagation* que Rumelhart, Hinton et Williams admettent dès l'article de 1986 : « L'inconvénient le plus évident de la procédure d'apprentissage est que l'espace des erreurs peut avoir des minima locaux, de sorte qu'il n'est pas garanti que la descente de gradient trouve un minimum global[185]. »

Afin d'éclaircir ces propos, revenons sur le fonctionnement de l'algorithme de *back-propagation.* Le réseau de neurones est d'abord paramétré au hasard. Avec ces paramètres, le réseau de neurones classe les données d'entraînement, se trompant dans certains cas, réussissant

181 Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason,* Cambridge, MIT Press, 1992 (3è édition).

182 Steven Pinker et Alan Prince, « On Language and Connectionism. Analysis of a Parallel Distributed Processing Model of Language Acquisition », *Cognition,* vol. 28, 1988, p. 73-193.

183 Jerry Fodor et Zenon Pylyshyn, « Connectionism and Cognitive Architecture: A Critical Analysis », *Cognition,* 28, 1988, p. 3-71.

184 Daniel Andler, *op. cit.*, p. 114-115.

185 « The most obvious drawback of the learning procedure is that the error-surface may contain local minima so that gradient descent is not guaranteed to find a global minimum. », Rumelhart, Hinton et Williams, *op. cit.*

dans d'autres. L'écart entre les réponses données par le réseau de neurones et les bonnes réponses donne un taux d'erreur. De petites variations des paramètres sont alors introduites. Si elles font baisser le taux d'erreur, elles sont conservées. Ainsi, pas à pas, le taux d'erreur est amené à un minimum : aucune petite variation ne permet de le faire baisser davantage. En d'autres termes, l'algorithme « cherche » les paramètres (poids et biais) pour lesquels le taux d'erreur (l'écart entre les prédictions de l'algorithme et les « bonnes réponses » fournies par la base de données d'entraînement) est le plus faible : l'algorithme *minimise* la fonction du taux d'erreur. Or, la forme de la fonction à approximer est inconnue. Elle pourrait avoir *plusieurs* minima et faire partie des fonctions dites « non convexes ». Si cela est le cas, comment savoir si le minimum atteint par l'algorithme est le bon ?

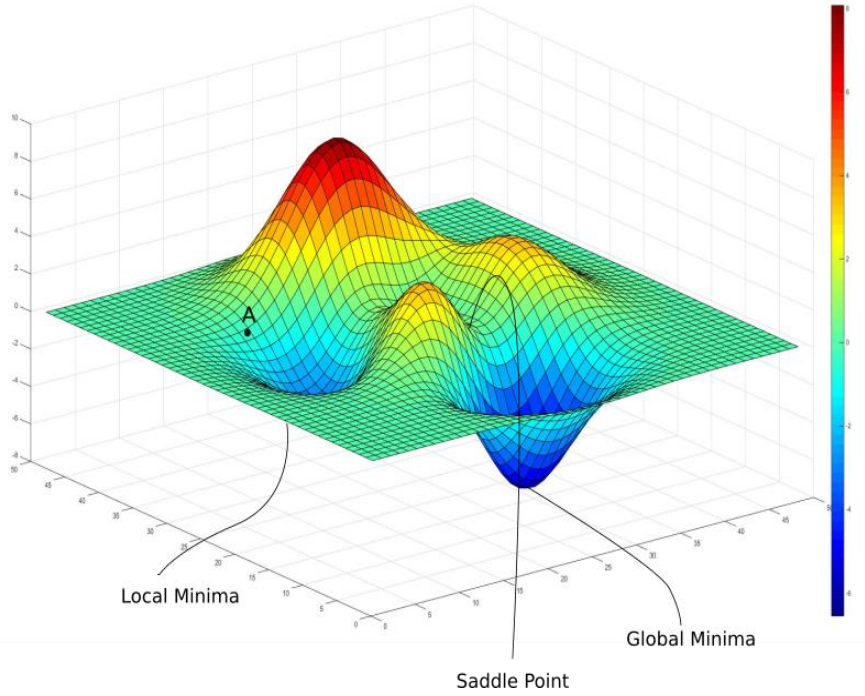


Illustration d'une fonction ayant plusieurs minima, Ayoosh Kathuria, « Intro to optimization in deep learning: Gradient Descent »,
https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

L'algorithme peut être comparé à un marcheur de montagne qui chercherait le point le moins élevé (minimum global). Le marcheur part d'un point au hasard et se dirige toujours vers le bas. Une fois le marcheur arrivé au fond d'un vallon (un minimum), il s'arrête puisqu'il ne « voit » pas vers où descendre. Mais comment savoir s'il s'agit bien du vallon le plus bas de toute la vallée (minimum global) ? Le « problème de la convexité » désigne le fait qu'une fois la

recherche effectuée, rien ne garantit que le minimum trouvé soit global. Or, on ne sait pas si la fonction est « non convexe ». Il se peut qu'elle ait un minimum global et plusieurs minima locaux.

> En raison de leur non-linéarité constitutive, les réseaux de neurones ne peuvent pas garantir que lors de la phase d'optimisation de la fonction de perte, le minimum global ait été trouvé ; il se peut très bien qu'elle converge vers un minimum local ou un plateau[186].

Cela a pu amener à penser que l'algorithme ne *devrait pas marcher* puisque de toute évidence il a plus de chances de tomber sur des minima locaux que sur le minimum global. Pour Rumelhart et ses collègues, si en théorie le système peut tomber dans des minima locaux, *en pratique* il apporte une solution « acceptable » : si le minimum trouvé n'est pas à coup sûr le minimum global, il en semble proche. Devant l'objection théorique, ils ne proposent qu'une réponse empirique : il se trouve que, dans la majorité des cas, cela fonctionne : « renouveler l'expérience avec de multiples tâches montre que le réseau n'est que très rarement bloqué dans un minimum local qui serait significativement plus mauvais que le minimum global[187]. » Et cela fonctionne d'autant mieux si les réseaux sont de grande taille[188]. Minsky et Papert réagissent avec virulence : dire que le système fonctionne « dans la plupart des cas », revient à manquer de sérieux en « prétend[ant] que le problème n'existe pas[189] ». Pour eux, les travaux récents ne sont pas assez solides pour rouvrir la controverse de 1969 : « La situation ne semble pas avoir beaucoup changé – nous n'avons pas vu de publication connexionniste contemporaine qui apporte de nouvelle lumière théorique[190]. »

En plus de jeter le discrédit sur les réseaux de neurones, le problème de la convexité a pour effet de décourager leur utilisation en tant qu'outil industriel. Leurs performances ne disposant d'aucune garantie théorique, on leur préfère d'autres techniques concurrentes, notamment les *support vector machines* (SVM). Il faut plusieurs dizaines d'années pour que le

186 Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones, L'invention des machines inductives et la controverse de l'intelligence artificielle », in *Réseaux* 2018/5 (n° 211), p. 173 à 220.

187 « However, experience with many tasks shows that the network very rarely get stuck in poor local minima that are significantly worse than the global minimum. » Rumelhart, Hinton et Williams, *op. cit.*

188 « We have only encountered this undesirable behaviour in networks that have just enough connections to perform the task. Adding a few more connections creates extra dimensions in weight-space and these dimensions provide path around the barriers that create poor local minima in the lower dimensional subspaces. » Rumelhart, Hinton et Williams, *op. cit.*

189 « Such pronouncements are not merely technically wrong; more significantly, the pretense that problems do not exist can deflect us from valuable insights that could come from examining things more carefully. » Marvin Minsky et Seymour Papert, cités par Olazaran, *op. cit.*, p. 648.

190 « The situation seems not to have changed much - we have seen no contemporary connectionist publication that casts much new theoretical light on the situation. » *Ibid.*

gain d'efficacité apporté par les réseaux de neurones finisse par prévaloir sur le manque de garanties théoriques, grâce aux efforts assidus de leurs partisans.

> 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)[191].

Pour le groupe de la « conspiration des neurones », dans la mesure où il est établi que, sur certaines tâches, les réseaux de neurones sont plus performants que les SVM, autant les utiliser – la théorie viendra après. Pour Yann LeCun, « 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[192] ». À partir de l'exemple de la machine à vapeur, inventée et adoptée avant la théorie thermodynamique, il défend l'idée que lorsqu'il y a innovation, il est courant que *la pratique précède la théorie* – comme si la raison était *légèrement dépassée* par sa propre création.

Il faudra de nombreuses années pour ces arguments fassent mouche et que les performances grandissantes des réseaux de neurones arrivent à convaincre la communauté de la recherche en intelligence artificielle, puis celle des informaticiens en général. Ainsi, alors que les algorithmes de rétropropagation sont inventés dans les années quatre-vingt et font leurs preuves dès les années 1990, les partisans des réseaux de neurones restent cantonnés « aux marges de la communauté de l'apprentissage artificiel[193] » jusque dans les années 2010.

1.2.13. La « révolution » du *deep learning* (les années 2010)

Trois éléments vont permettre aux réseaux de neurones de s'imposer dans de nombreux domaines de l'informatique et de gagner une nouvelle crédibilité dans le champ de l'intelligence

191 Cardon et al. *Ibid.*

192 Lors d'une conférence en 2007, « 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). » Cardon et al., *Ibid.*

193 Cardon et al., *Ibid.*

artificielle. Un premier est la constitution d'importantes bases de données d'exemples correctement labellisés[194] permettant d'entraîner les réseaux. ImageNet, mise à disposition des chercheurs par Fei-Fei Li en 2009, offre ainsi 3,2 millions d'images illustrant des dizaines de milliers de catégories[195].

Un deuxième est l'accroissement exponentiel de la puissance du *hardware* disponible, permettant d'entraîner les réseaux de neurones sur ces gigantesques bases de données. Plutôt que d'utiliser des processeurs classiques (CPU), certains chercheurs ont l'idée de recourir aux processeurs de cartes graphiques (GPU), dont la puissance se développe exponentiellement sous l'impulsion de l'industrie du jeu vidéo, et dont l'architecture parallèle est plus propice à l'entraînement d'un réseau de neurones[196].

Enfin, un troisième est la démultiplication d'ajustements, d'innovations et d'améliorations apportés aux différents réseaux de neurones. Une communauté de chercheurs travaille à améliorer les performances des programmes dans l'accomplissement de tâches précises (reconnaissance d'image ou de son, synthèse vocale, etc.). La communauté se désigne désormais comme celle du *deep learning*, une terminologie déjà ancienne, mais qui a l'avantage de bien marquer la différence avec les réseaux de neurones (dits *shallow*) de l'époque de Rosenblatt[197].

Par rapport aux réseaux de neurones des années quatre-vingt qui avaient peu de couches (trois ou quatre), un seul type de neurones et des connexions entre chaque neurone d'une couche à l'autre, les réseaux dits de *deep learning* peuvent avoir jusqu'à des centaines de couches, ils

194 On dit d'un exemple qu'il est « labellisé » lorsqu'est indiqué ce que la machine doit répondre. L'assemblage et l'annotation minutieuse d'une base de données telle qu'ImageNet est un travail considérable. Etant donné l'ampleur de la tâche, cela n'aurait sans doute pas été possible sans l'émergence des plate-formes de *crowdsourcing,* et en l'occurence Amazon Mechanical Turk. ImageNet est restée une référence pour la vision artificielle mais fait l'objet de critiques à cause des labels caricaturaux, racistes ou misogynes, qui ont été appliqués aux images d'êtres humains. Kate Crawford et Trevor Paglen, « Excavating AI, The Politics of Images in Machine Learning Training Sets », https://excavating.ai/, page consultée le 20 mars 2021.

195 Fei-Fei Li, Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, « ImageNet: A Large-Scale Hierarchical Image Database », *Conference on Computer Vision and Pattern Recognition CVPR,* 2009.

196 Kumar Chellapilla, Sidd Puri, Patrice Simard, « High Performance Convolutional Neural Networks for Document Processing », *Tenth International Workshop on Frontiers in Handwriting Recognition*, Université de Rennes 1, Oct 2006, La Baule (France). Inria-00112631. Voir également Dan Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, Jürgen Schmidhuber, « Flexible, High Performance Convolutional Neural Networks for Image Classification », *International Joint Conference on Artificial Intelligence IJCAI*, 2011, 1237-1242.

197 Pour une histoire détaillée de la fabrication du *deep learning,* et sur la différence entre *shallow* et *deep learning,* voir Jurgen Schmidhuber, « Deep Learning in Neural Networks: An Overview », *Neural Networks,* volume 61, janvier 2015, p. 85-117.

sont hétérogènes[198], leurs connections sont réduites[199], et évitent l'*overfitting*[200] grâce à différentes méthodes comme l'ajout de bruit, la variation des images, le *dropout* (désactivation de certaines parties du réseau pour que son pouvoir de discrimination ne repose pas exclusivement sur certaines zones), ou des modifications de la fonction d'erreur favorisant les solutions plus simples[201] .

À la faveur de ces trois éléments, les réseaux de neurones parviennent à une efficacité sans précédent. En 2011 et 2012, Dan Ciresan et ses collègues remportent une série de compétitions de reconnaissance d'image[202]. Les réseaux convolutionnels tournant sur des GPUs dépassent les autres méthodes pour reconnaître des caractères chinois, des panneaux de signalisation, ou des images médicales. En 2012, à l'occasion d'une compétition basée sur ImageNet, la victoire écrasante de l'équipe de Hinton[203] bouleverse la communauté de la vision artificielle en montrant à quel point les autres approches sont devenues obsolètes.

Le raz-de-marée qui s'opère dans le champ de la reconnaissance d'image ne laisse pas les autres domaines en reste.

> 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[204].

En 2015, Hinton, LeCun et Bengio font le point dans *Nature*[205] : les différents algorithmes de *deep learning* se montrent d'une supériorité nette pour les tâches de traitement de l'image, de

198 Ainsi, au lieu d'avoir un seul type de neurones, les réseaux convolutionnels en ont trois : les *convolutional nodes*, les *rectified linear units* (ReLU), et les *non-linear downsamplers* qui compilent les caractéristiques extraites par les *convolutional nodes* et transmises par les ReLU.

199 Le terme d'usage est « sparse connectivity ». L'idée est de réduire le nombre de paramètres à calibrer.

200 Il y a *overfitting* lorsqu'un algorithme d'apprentissage « décalque » trop les particularités des exemples de la base d'entraînement. Il n'est alors pas capable de « généraliser », c'est-à-dire de s'abstraire de ces particularités pour traiter des cas différents.

201 Nous restituons les quatre caractéristiques données par Cameron Buckner, « Deep Learning: A Philosophical Introduction », *Philosophy Compass*, 14, 2019.

202 Dan Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, Jürgen Schmidhuber, « Flexible, High Performance Convolutional Neural Networks for Image Classification », *International Joint Conference on Artificial Intelligence IJCAI*, 2011, p. 1237-1242.

203 Krizhevsky, Alex, Ilya Sutskever et Geoffrey Hinton « ImageNet Classification with Deep Convolutional Neural Networks » *NIPS 2012: Neural Information Processing Systems*, Lake Tahoe, Nevada, 2012.

204 Cardon et al., *op. cit.*

205 LeCun, Yann, Yoshua Bengio, Geoffrey Hinton, « Deep learning », *Nature*, vol. 521, n° 7553, 2015.

la vidéo, du son et de la parole (*deep convolutional nets*), ainsi que la manipulation de texte (*recurrent nets*, plus adaptés aux données séquentielles).

Les algorithmes de *deep learning* ne s'appliquent pas à toutes les tâches, mais pour celles où leur supériorité est manifeste, le basculement se fait très rapidement. Les informaticiens délaissent les approches traditionnelles pour adopter le *deep learning*. En 2012, pour la compétition ImageNet, seule l'équipe de Hinton utilisait un réseau convolutionnel tournant sur des GPUs. En 2014, la plupart des compétiteurs ont adopté la même architecture et cherchent à y apporter leurs propres améliorations.

En 2016, avec la médiatisation de la victoire d'AlphaGo contre Lee Sedol[206], le grand public découvre l'incroyable efficacité du *deep learning.* Entre temps, ce dernier a trouvé de nombreuses applications industrielles. Il est utilisé quotidiennement par les *smartphones* et enceintes connectées dotés de reconnaissance vocale, ainsi que dans l'industrie, pour des tâches aussi variées que la détection de défauts, l'optimisation sous contrainte, la maintenance prédictive, etc.

Ce sont les géants du numérique qui, les premiers, adoptent le *deep learning* et investissent massivement dans son développement. Dès 2013, Google embauche le « pionnier » Geoffrey Hinton, tandis que Facebook recrute Yann LeCun[207]. Les moyens colossaux qui en découlent, combinés à la popularité du *deep learning,* ainsi qu'à son efficacité indéniable pour certaines tâches, entraînent un afflux considérable d'informaticiens – nouveaux arrivants ou anciens reconvertis. La « communauté » du *deep learning*, et plus généralement celle du *machine learning*, se démultiplie, comme en témoigne l'évolution exponentielle du public de ses principales conférences. Alors que l'International Conference on Machine Learning acceptait en moyenne 200 contributions, elle en compte 1088 en 2020 (pour 4990 propositions[208]). La pression est encore plus forte pour la conférence *Neural Information Processing System* (NeurIPS) : habitués à recevoir quelques centaines de propositions, les organisateurs en reçoivent plus de 1500 en 2014, plus de 3000 en 2017, et près de 9500 en 2020[209].

206 Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Hui Fan, Laurent Sifre, George van den Driessche, Thore Graepel et Demis Hassabis, « Mastering the game of Go without human knowledge », Nature, vol. 550, no 7676, 19 octobre 2017, p. 354–359.

207 Cade Metz, *Genius Makers, The Mavericks Who Brought AI to Google, Facebook, and the World*, New-York, Penguin Random House, 2021.

208 Source, *Microsoft Research*, https://www.microsoft.com/en-us/research/project/academic/articles/icml-conference-analytics/ page consultée le 20 mars 2021.

209 Diego Charrez, « NeurIPS 2019 Stats », https://medium.com/@dcharrezt/neurips-2019-stats-c91346d31c8f

Par rapport aux réseaux neuronaux simples (*shallow network*), dont les propriétés sont bien comprises, les différentes branches du *deep learning* ouvrent à une complexité qui dépasse l'état des connaissances[210]. Mais cela n'a plus la même importance que lors des controverses des années quatre-vingt : « 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[211]. »

1.2.14. Chronologie succincte

1936 Alan Turing propose la notion fondatrice pour l'informatique de **machine universelle.**
1938 Dans sa thèse, Alan Turing évoque l'intuition et se refuse à la comparer à une machine
1943 McCulloch et Pitts démontrent qu'un **réseau de neurones** peut réaliser des opérations logiques et fonctionner comme une machine universelle de Turing.
1946 présentation de l'ENIAC, un des premiers **ordinateurs programmables.**
1950 Alan Turing introduit l'idée d'un **test de l'intelligence** appliqué aux machines.
1955-56 McCarthy, Minsky, Rochester et Shannon publient un appel à projet pour un séminaire d'été à l'université de Dartmouth. Ils formulent la conjecture qui fondent leurs travaux : « tous les aspects de l'apprentissage ou toute autre caractéristique de l'intelligence peuvent en principe être décrits d'une manière si précise qu'une machine peut être fabriquée pour les simuler ». C'est la première fois qu'est utilisé le terme d'« **intelligence artificielle** ».
1957 Frank Rosenblatt fabrique le **Perceptron**, un réseau de neurones capable de reconnaître des formes.
1957-1970 L'intelligence artificielle prend le statut de **discipline scientifique** : se créent des laboratoires (MIT, Stanford…) qui reçoivent des financements des gouvernements et en particulier de l'armée. Cela permet de nombreuses inventions (jeux d'échecs, chatbot, démonstration mathématique) qui sont abondamment relayées par la presse. Les chercheurs promettent la construction de machines intelligentes en quelques années.

210 Stéphane Mallat, « Understanding deep convolutional networks », *Philosophical Transactions of the Royal Society* A, Volume 374, Issue 2065, 13 avril 2016.
211 Cardon et al., *op. cit.*

1969-1970 Plusieurs publications critiquent l'enthousiasme du public, les promesses excessives des chercheurs et mettent en avant les limites théoriques des machines. Minksy et Papert s'attaquent en particulier au Perceptron de Rosenblatt. Cela cause un premier « **hiver de l'IA » :** le grand public se désintéresse de l'IA et les financements sont restreints, au seul profit de l'école symbolique.

Années 1980 Retour en grâce du projet d'IA avec les **systèmes experts** qui reproduisent le comportement d'experts sous la forme de règles d'action. Les premières *startups* d'intelligence artificielle apparaissent. Des financements privés s'ajoutent aux financements publics. Après quelques années de succès, le coût élevé des programmes et les limites des moteurs de règles entraînent la déception des financeurs et un **deuxième hiver de l'IA.**

1986 Les réseaux de neurones, discrédités depuis 1969, sont réhabilités grâce aux travaux de Rumelhart, Hinton et Williams qui inventent un nouvel algorithme dit de **backpropagation**.

1989 Yann LeCun invente un type de réseau de neurones capable de reconnaissance d'image, les **réseaux convolutionnels,** ils sont utilisés dès les années 1990 par la poste américaine pour la lecture automatique de chèques.

2006 Hinton *et al.* proposent le terme de **deep learning** pour désigner les réseaux de neurones qui comportent de nombreuses couches.

2012 Hinton *et al.* remportent une compétition de vision artificielle et déclenchent une prise de conscience de la communauté informatique qui adopte massivement le *deep learning*.

Années 2010 Différentes familles d'algorithmes de deep learning s'imposent progressivement dans une grande variété de domaines : traitement du son, traduction, traitement du langage naturel, jeu de go… Les géants du numérique créent leurs laboratoires dédiés et intègrent ces technologies à leurs produits.

1.3. L’apprentissage profond (deep learning)

1.3.1. Réseaux de neurones, *machine learning* et intelligence artificielle

Bien que les réseaux de neurones aient joué un rôle privilégié dans l’histoire de l’intelligence artificielle, ils ne sont **qu’une piste de recherche parmi d’autres** dans la myriade d’inventions qu’a connu l’histoire de l’intelligence artificielle. Il faut aussi prendre en considération les programmes du courant symbolique (moteurs de règles, knowledge graph…), ainsi que tout ceux qui font partie de l’apprentissage automatique (aussi appelé *machine learning*) sans faire partie de la famille des réseaux de neurones : algorithme de régression logistique, machines à vecteur de support, arbres de décision et forêts aléatoires, K-moyennes, programmation génétique[212]…

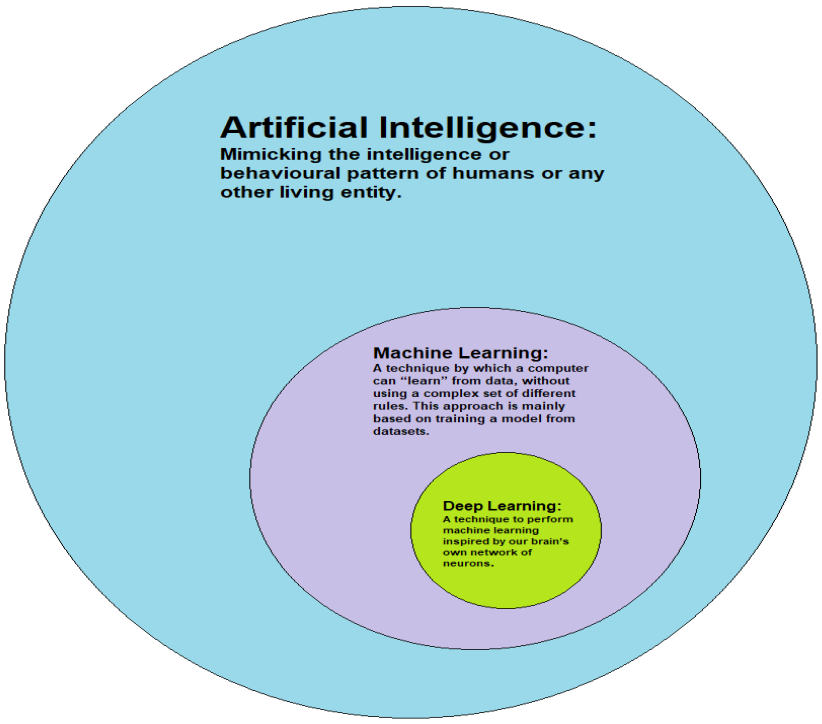


Image « How deep learning is a subset of machine learning and how machine learning is a subset of artificial intelligence (AI) », Avimanyu Bandyopadhyay, Hands On GPU Computing with Python, *Packt Publishing, 2019.*

212 Pour un recensement des différentes méthodes d’apprentissage, voir la thèse d’Antoine Mazières : *Cartographie de l’apprentissage artificiel et de ses algorithmes,* thèse soutenue à l’Université Paris Diderot, sous la direction de Jean-Philippe Cointet, 2016.

Les concepteurs des réseaux de neurones se rattachent explicitement au projet d'intelligence artificielle. Par contre, dans les autres domaines du *machine learning*, les chercheurs n'ont pas toujours manifesté d'intérêt pour la fabrication de machines pensantes. Par exemple, les travaux de Vladimir Vapnik, co-inventeur des *support-vector machines* (ou SVM) mentionnées précédemmment, ont été présentés dans le cadre de ses travaux sur « l'apprentissage statistique », sans référence à l'explication ou l'imitation de l'intelligence humaine. Avant la fabrication de machines intelligentes, c'est l'efficacité dans la résolution de problèmes donnés qui guide la majorité des recherches.

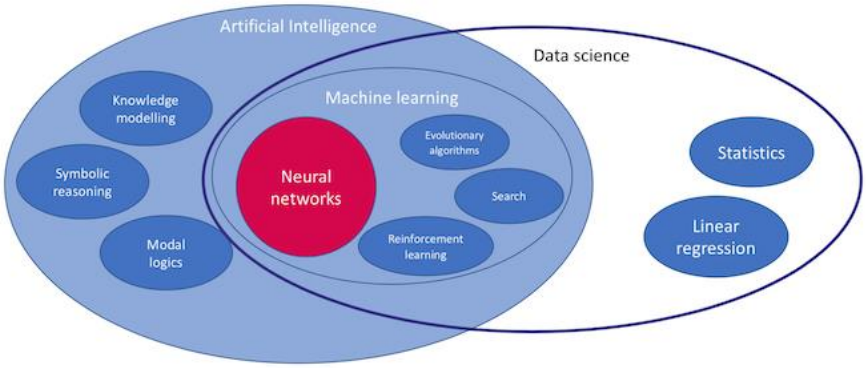


Image : Sieuwert van Otterloo, « AI, Machine learning and neural networks explained », ICT Institute, 27 juillet 2020, https://ictinstitute.nl/ai-machine-learning-and-neural-networks-explained/, page consultée le 20 septembre 2020.

Aujourd'hui, l'engouement pour l'intelligence artificielle doit moins à la pertinence théorique des algorithmes qu'à leur efficacité pratique. Leur portée scientifique peut se discuter, mais leur succès industriel est indéniable. Les algorithmes sont utilisés tous les jours, pour un nombre croissant de tâches et dans des contextes variés. S'ils reçoivent autant de financements et d'attention médiatique, ce n'est pas en tant que modèle de la cognition, mais parce que ce sont des outils efficaces. Lorsque les équipes de Yann LeCun au FAIR améliorent les outils de reconnaissance d'image, la question qui est posée n'est pas « s'agit-il d'un modèle pertinent de la perception ? » mais « sommes-nous capables d'associer le bon texte aux images postées sur Facebook, de manière à pouvoir les classer et les intégrer à notre machine publicitaire ? ».

Avant d'étudier la portée épistémologique de ces modèles, il convient donc d'exposer brièvement en quoi consiste cet usage industriel.

1.3.2. Le *machine learning* en application

Une fois conçus par des chercheurs (affiliés ou non au champ de l'intelligence artificielle), les algorithmes sont récupérés, modifiés et mis en application par des praticiens qui se réclament moins de l'intelligence artificielle que du *machine learning* ou de la *data science*. Dans le cas d'un apprentissage supervisé[213], nous restituons ici leur travail en huit étapes, en nous inspirant des six étapes proposées par Maël Pegny et Mohamed Issam Ibnouhsein[214], ainsi que par celles proposées par Jérémie Jakubowicz[215] :

1. Choix et formalisation du problème. Les *data scientists* sont sollicités pour une varieté de problèmes tels que :

- **optimisation** : trouver le jeu de paramètres qui maximise ou minimise une production. Par exemple, paramétrer une usine pour maximiser sa production ou un centre de données pour minimiser sa consommation d'électricité ;

- **prédiction** : prédire les valeurs futures d'une variable à partir de son historique, par exemple la température ou le nombre de passagers du métro ;

- **classification** : attribuer un label à un objet à partir de certaines caractéristiques, par exemple reconnaître un objet ou une personne dans une image ou détecter des opérations bancaires suspectes ;

- **association** : associer deux objets sur la base de critères définis, par exemple pour un moteur de recherche (associer une demande et une réponse), un site de rencontre ou un moteur de recommandation[216].

Pour illustrer notre propos, nous prenons **le cas d'un algorithme de détection de fraude à la réclamation dans la vente en ligne**[217]. Il s'agit d'un problème de classification puisqu'il faut attribuer un label à chaque dossier de réclamation : suspect ou non suspect.

213 L'apprentissage est dit « supervisé » lorsqu'il se fait à partir d'exemples labellisés. Il se distingue de formes d'apprentissage non supervisé, comme l'apprentissage par renforcement (le programme se calibre à partir du *feedback* de la situation) ou encore les algorithmes de *clustering* qui font émerger des catégories en regroupant les données par similarité. Nous ne présentons que le cas de l'apprentissage supervisé car celui-ci représente aujourd'hui la majorité des usages.

214 Maël Pegny et Mohamed Issam Ibnouhsein, « Quelle transparence pour les algorithmes d'apprentissage machine ? » 2018, hal-01791021, p. 17-18.

215 Lors d'une conférence donnée à l'ENS le 12 novembre 2019.

216 Pour une présentation plus détaillée des algorithmes en général, et de leur fonction de recommandation en particulier, voir Dominique Cardon, *A quoi rêvent les algorithmes*, Paris, Seuil, 2015.

217 Nous empruntons cet exemple à Jérémie Jakubowicz (conférence du 12 novembre 2019).

2. Collecte de données. Une fois le problème défini, le *data scientist* récupère les données nécessaires, comme l'historique de trafic, les profils d'utilisateurs ou la liste de réclamations. Pour le cas de la fraude à la réclamation, il s'agit d'un tableau contenant une ligne par réclamation et une cinquantaine de colonnes avec le nom de la personne qui réclame, son adresse, le point de départ du colis, le point d'arrivée, le circuit, le poids, la somme demandée…

Dans l'exemple qui nous occupe, la base de données utilisée provient d'enquêtes déjà réalisées et contient le label « réclamation frauduleuse » ou « non frauduleuse ». Si l'information n'est pas déjà présente, il faut **labelliser les données** en fonction de ce que l'algorithme doit « apprendre ». Ce travail manuel peut être long et fastidieux[218] mais son importance est considérable. S'il est mal fait, l'algorithme risque de se calibrer sur des variables non pertinentes.

Plus généralement, la pertinence de l'algorithme dépend de la qualité de la base de données. Les exemples doivent être suffisamment variés pour que l'algorithme puisse en abstraire des caractéristiques assez générales pour s'appliquer à la plupart des cas. Par exemple, les algorithmes de détection de mélanomes sont défaillants face à des peaux non blanches si leur base d'entraînement n'en contient pas assez[219].

3. La **préparation des données ou *feature engineering*** est une étape très chronophage. Une fois les données récupérées, le *data scientist* les « nettoie » de façon à ce qu'elles puissent être traitées par l'algorithme. Il corrige les erreurs et enquête pour déceler et élucider d'éventuelles incohérences. Une variété de méthodes permet de **combler les données manquantes** : supprimer la ligne ou la colonne incriminée, remplacer la donnée manquante par une donnée vraisemblable comme la médiane des autres données ou la même donnée qu'une ligne similaire (« stratégie du masquage »), indiquer une valeur négative ou exagérément élevée pour mettre en relief la case vide (« stratégie de l'isolement »). En ce qui concerne les données qualitatives (adresses, type de transporteur…), elles sont assignées à des catégories ou transformées en données quantitatives, une seule donnée qualitative pouvant être encodée dans plusieurs nouvelles colonnes de données quantitatives.

Les données « nettoyées » sont également **« enrichies »**. Le *data scientist* fait l'hypothèse que certaines informations complémentaires pourraient être pertinentes et ajoute

218 Dominique Cardon et Antonio Casilli ont enquêté sur les « microtravailleurs » dont une des tâches est la labellisation de bases de données. Dominique Cardon et Antonio Casilli, *Qu'est-ce que le Digital Labor ?*, Bry-sur-Marne, INA, 2015. Antonio Casilli, *En attendant les robots*, Paris, Éditions du Seuil, 2019.

219 Adamson, Smith « Machine Learning and Health Care Disparities in Dermatology » *JAMA Dermatology,* 2018; 154 (11) : 1247–1248.

des colonnes provenant d'autres bases de données, par exemple des données statistiques de l'INSEE sur la zone d'où provient la réclamation.

A la suite du *feature engineering*, le tableau de notre exemple passe de cinquante à deux cents colonnes.

4. Ensuite, le *data scientist* procède à **la sélection ou la conception de l'algorithme et des hyperparamètres**, c'est-à-dire les paramètres qui calibrent l'algorithme avant l'apprentissage. Dans le cas d'un réseau de neurones, cela correspond au nombre de couches et au nombre de neurones par couches.

- pour une première famille de cas, le problème est courant et l'algorithme correspondant est connu. Par exemple, il est commun de recourir aux réseaux de neurones pour tout ce qui a trait à la reconnaissance d'images. Dans ce cas, le *data scientist* n'a pas besoin de concevoir d'algorithme. Il lui suffit de **reprendre des algorithmes préexistants[220] et d'en choisir les hyperparamètres**,

- pour une deuxième famille de cas, le problème est courant mais plusieurs algorithmes classiques peuvent s'appliquer. Le *data scientist* **essaye les différents algorithmes et conserve celui qui produit les meilleurs résultats**.

- pour une troisième famille de cas, le problème est atypique et il **conçoit et code un modèle statistique *ad hoc***.

En d'autres termes, si dans certains cas le praticien a une bonne compréhension du problème et de la solution qu'il apporte, il arrive également qu'il procède **par essai-erreur**, en particulier chez les praticiens du *deep learning* lors de la sélection des hyperparamètres. Cette sélection s'effectue à tâtons, **de manière empirique**, sans qu'une théorie vienne justifier le choix. Bien que les praticiens partagent leur expérience et travaillent à l'élaboration d'un savoir commun, il leur faut passer beaucoup de temps à tester des configurations variées. Avec l'expérience, ce temps de tâtonnement diminue, les praticiens développant une familiarité avec les outils, qui se traduit par une capacité à éliminer d'emblée certaines configurations non pertinentes. Ce fonctionnement à l'aveugle a valu à leur pratique le qualificatif de « magie noire » (*black magic*)[221].

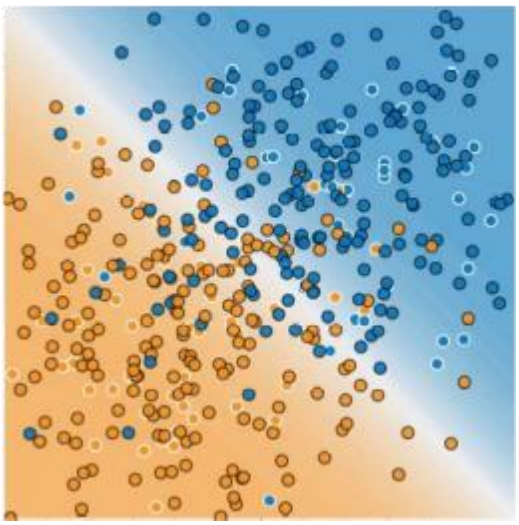
220 Un grand nombre d'algorithmes sont disponibles via des librairies *open source* comme Scikit-Learn, TensorFlow, Keras, etc.

221 Kanav Anand, Ziqi Wang, Marco Loog, Jan van Gemert, « Black Magic in Deep Learning: How Human Skill Impacts Network Training », « Black Magic in Deep Learning », arXiv:2008.05981, 2020.

Enfin, il est fréquent que les outils conçus **combinent plusieurs algorithmes** qui se répartissent le travail : par exemple, un réseau de neurones annote des images de façon à nourrir une base de données utilisée par un autre algorithme.

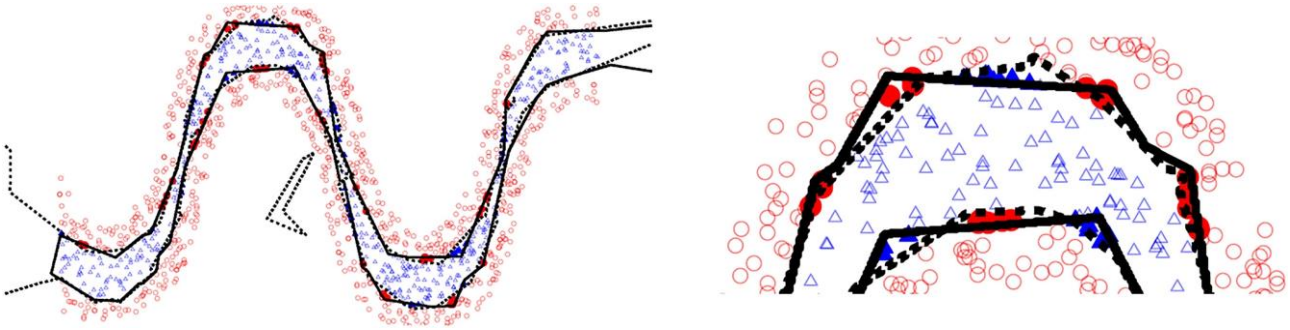
5. Lors de l'**apprentissage**, l'algorithme se calibre à partir des données préparées par le *data scientist*. Il effectue une procédure l'amenant à identifier les paramètres les plus pertinents pour réaliser sa tâche (classification, prédiction…).

Dans le cas des dossiers de réclamation, il s'agit de **classer** les différents dossiers, c'est-à-dire identifier au sein des données les critères permettant de distinguer les dossiers suspects des autres. Chaque ligne du tableau de données peut être vue comme un point dans un espace de dimension deux cents. Chaque point a été préalablement qualifié comme suspicieux (en jaune) ou non (en bleu). L'ensemble des lignes du tableau forme un nuage de points. L'objectif de l'algorithme est de **tracer une frontière** permettant de séparer les points jaunes des points bleus. Selon le type d'algorithme choisi, la forme de la frontière sera différente. Ainsi les algorithmes de régression linéaires tracent un plan, tandis que les algorithmes dits de forêts aléatoires ou de *boosting* forment un quadrillage.



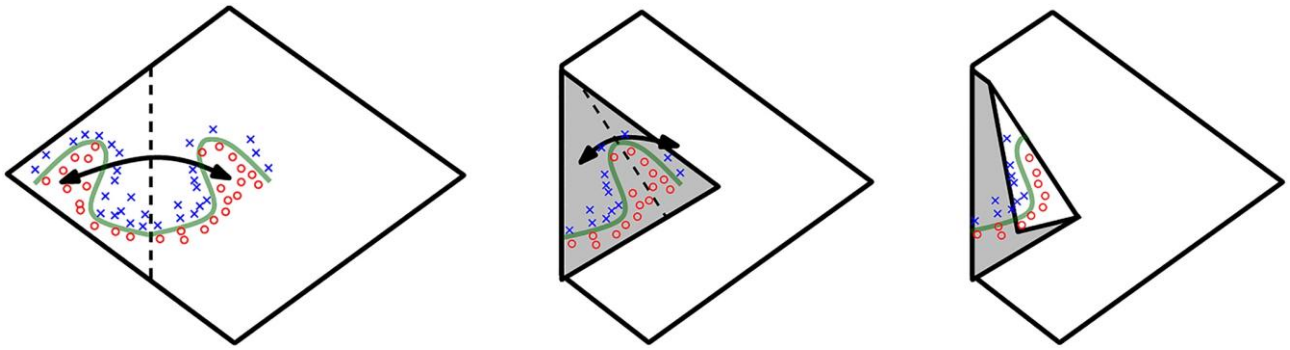
Ici, un algorithme de régression linéaire trace un plan permettant de délimiter la zone contenant le plus de points bleus et celle contenant le plus de points jaunes. Image : Machine Learning Crash Course, https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data *page consultée le 20 septembre 2020*.

Dans le cas des réseaux de neurones, le passage du *shallow* au *deep learning* a permis d'augmenter la précision de la classification :



Comparaison entre l'opération de classification (trouver une fonction qui sépare les points rouges des bleus) réalisée par un réseau de neurones classique ou shallow *(ligne continue) et un réseau* deep *(ligne discontinue). Les couleurs accentuées signalent les erreurs faites par le réseau classique. Image : Montufar et al. « On the Number of Linear Regions of Deep Neural Networks », arXiv:1402.1869, 2014.*

Une des explications données à l'efficacité des réseaux de neurones profonds est qu'ils identifient des symétries dans l'espace des points, ce qui leur permet de le « plier » et d'aligner les zones similaires pour y appliquer la même classification. Ainsi, chaque « pliage » successif permet de réutiliser une classification linéaire déjà opérée. Le processus d'apprentissage revient à identifier *quels pliages de l'espace des points exploitent des symétries et permettent de réduire la fonction d'erreur à moindre coût.*

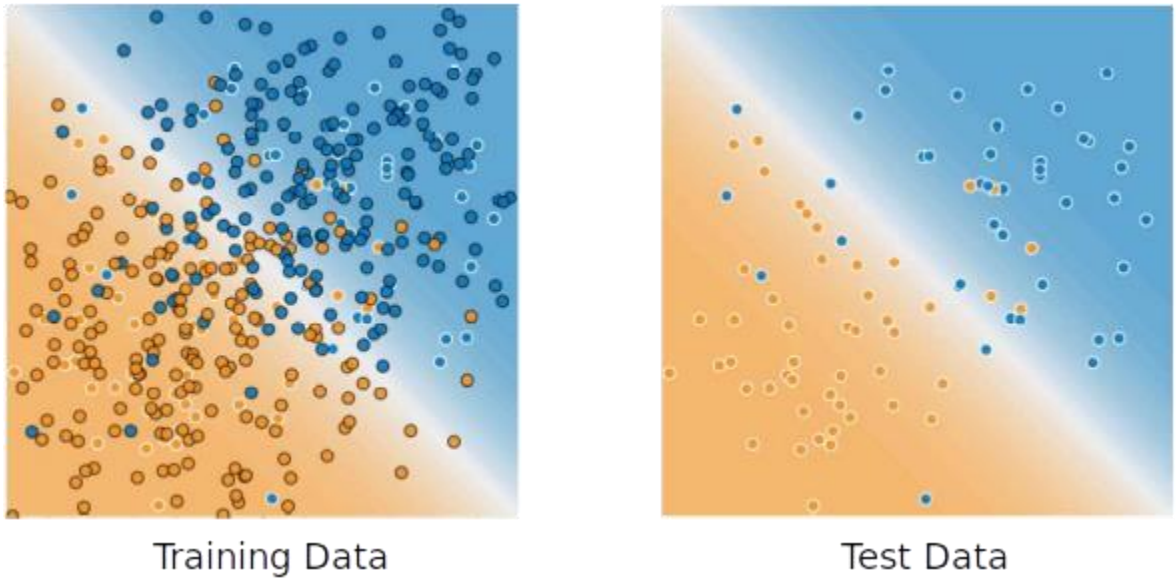


« Pliage » de l'espace des points permettant d'exploiter des symétries en appliquant les mêmes délimitations dans plusieurs zones. Image : Montufar et al. « On the Number of Linear Regions of Deep Neural Networks », arXiv:1402.1869, 2014.

Pour finir, précisons que la finesse de la frontière tracée ne doit pas être excessive. Si elle décalque trop précisément le nuage de points, elle intègre certaines particularités des exemples qui ne sont pas des caractéristiques pertinentes pour l'opération de classification. Il y

a alors *overfitting,* l'algorithme n'est pas capable de traiter de nouveaux cas – il ne sait pas « généraliser ».

5. Seule une partie des données disponibles est utilisée pour l'apprentissage (base de données dite « d'entraînement » ou *training set*), le reste est mis de côté pour **tester la performance de l'algorithme** (base de données de test ou *test set*). Le *data scientist* a gardé en réserve un certain nombre de points dont il connaît la couleur et qu'il soumet à l'algorithme pour voir si la frontière tracée permet de bien les qualifier – si les points rouges de test se retrouvent bien dans la zone des points rouges. Le ratio de points correctement classés donne une mesure de la performance de l'algorithme. C'est cet indicateur que le *data scientist* cherche à améliorer en essayant différents types d'algorithmes et différents hyperparamètres.



Grâce aux données d'entraînement, l'algorithme a « appris » à délimiter l'espace des points en une zone bleue et une zone jaune (image de gauche). Avec les nouveaux points fournis par les données de test, on vérifie que cette délimitation est pertinente (le moins de points bleus dans la zone jaune et vice-versa – image de droite). Image : Machine Learning Crash Course, https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data , *page consultée le 20 septembre 2020.*

Cette étape peut prendre plusieurs semaines et il est important de savoir la stopper : les *data scientists* tendent à rechercher toujours plus de précision, alors même qu'il est nécessaire que le modèle reste simple pour pouvoir traiter des données qui diffèrent des données d'entraînement.

6. Mise en production. Une fois que les performances obtenues sont jugées suffisantes par le *data scientist,* le modèle est « déployé », c'est-à-dire mis à disposition depuis un serveur de façon à pouvoir lui soumettre de nouveaux points à qualifier[222]. Dans le cas de la détection des fraudes, les employés du site de vente en ligne soumettent les dossiers de réclamation à l'algorithme pour qu'il les qualifie (suspect ou non).

7. Une fois le modèle déployé, le *data scientist* **teste sa pertinence en production** : est-ce que les dossiers qualifiés de frauduleux le sont effectivement ? D'une part, il importe de vérifier que l'algorithme est bien capable de « généraliser » – qu'il n'y a pas d'*overfitting* – c'est-à-dire qu'il s'applique à de nouvelles données. D'autre part, il est possible que la situation ait changé entre le stade de fabrication du modèle et sa mise en production, par exemple si les fraudeurs ont découvert de nouvelles stratégies (problème dit de la « dérive des concepts » ou *concept drift*).

8. La « dérive des concepts » et plus généralement l'afflux de nouvelles données amènent le *data scientist* à **mettre le modèle à jour** pour préserver et améliorer sa pertinence, à une fréquence qui varie selon la constance des phénomènes en cause : certains sont assez stables (maintenance prédictive pour des machines qui ne changent pas) ou très variables (comportement d'internautes ou de fraudeurs).

222 C'est une étape qui peut se révéler très chronophage en fonction de la complexité du système de circulation des données, complexité qui varie en fonction de leur fréquence de circulation (les données doivent-elles circuler en temps réel ou à intervalles régulier ?), de leur type (s'agit-il de texte, de son, d'images ? Les données sont-elles structurées ou non structurées ?) et de leur volume. C'est un métier à part entière que de mettre en place et dimensionner la bonne architecture logicielle et matérielle pour assurer le bon fonctionnement du système. Enfin, à cela peut s'ajouter la réalisation d'une interface utilisateur permettant d'intéragir avec le système.

1.3.3. La « boîte noire » du *deep learning*

De tous les algorithmes qui composent la boîte à outils des *data scientists,* le *deep learning* n'est pas le plus populaire. D'une part, de nombreux problèmes peuvent être résolus avec des algorithmes classiques plus simples, moins gourmands en calcul, et fournissant des résultats aussi bons, voire meilleurs. D'autre part, les réseaux de neurones souffrent d'un défaut d'explicabilité des résultats qui leur vaut le qualificatif de « boîte noire[223] ».

Dans le cas d'un réseau de neurones contenant, par exemple, 12 000 paramètres qui ont été ajustés par essai-erreur sur cinq millions d'exemples, il est difficile, lorsqu'il donne un *output,* de l'attribuer à des critères définis et à un cheminement logique clair. Une fois que les milliers de paramètres interdépendants sont arrivés à la combinaison la plus performante, il est impossible de restituer l'opération qu'effectue le réseau de neurone sous la forme d'une suite d'étapes explicites. On ne dispose que de la matrice des poids et des biais, mais pas d'une explication digne de ce nom. Ainsi, un algorithme d'apprentissage profond ne produit qu'une connaissance partielle : la procédure (ou heuristique) qui le constitue est connue, et il fournit un résultat pertinent (distinguer un chat d'un chien), mais les structures sous-jacentes des objets étudiés (« features » ou caractéristiques) sur lesquelles il s'appuye pour produire ce résultat sont inconnues[224].

Une telle réalisation sans compréhension n'aurait pas rebuté Turing, qui en perçoit l'éventualité dans son article de 1950 et admet à son test une machine « qui fonctionne, mais dont les modalités de fonctionnement ne peuvent être décrites de manière satisfaisante par ses constructeurs, parce qu'ils ont appliqué une méthode en grande partie expérimentale[225]. » Mais pour les usagers, un tel défaut d'explicabilité peut disqualifier l'algorithme. À quoi peut servir un outil de prédiction des pannes si celui-ci n'est pas capable d'expliquer sa décision, c'est-à-dire d'énoncer les causes probables de la panne ? L'opacité des réseaux de neurones peut également les rendre inconvenants. Un algorithme accompagnant l'attribution de crédits doit

223 Davide Castelvecchi, « Can we open the black box of AI ? », *Nature* 538, 20-23, 5 octobre 2016.

224 Pour plus de détails, Stéphane Mallat aborde ce point dans le cours au Collège de France du 30 janvier 2019, à la minute 18, https://www.college-de-france.fr/site/stephane-mallat/course-2019-01-30-09h30.htm, page consultée le 4 octobre 2019.

225 Alan Turing, « Les ordinateurs et l'intelligence », *op. cit.,* p. 139.

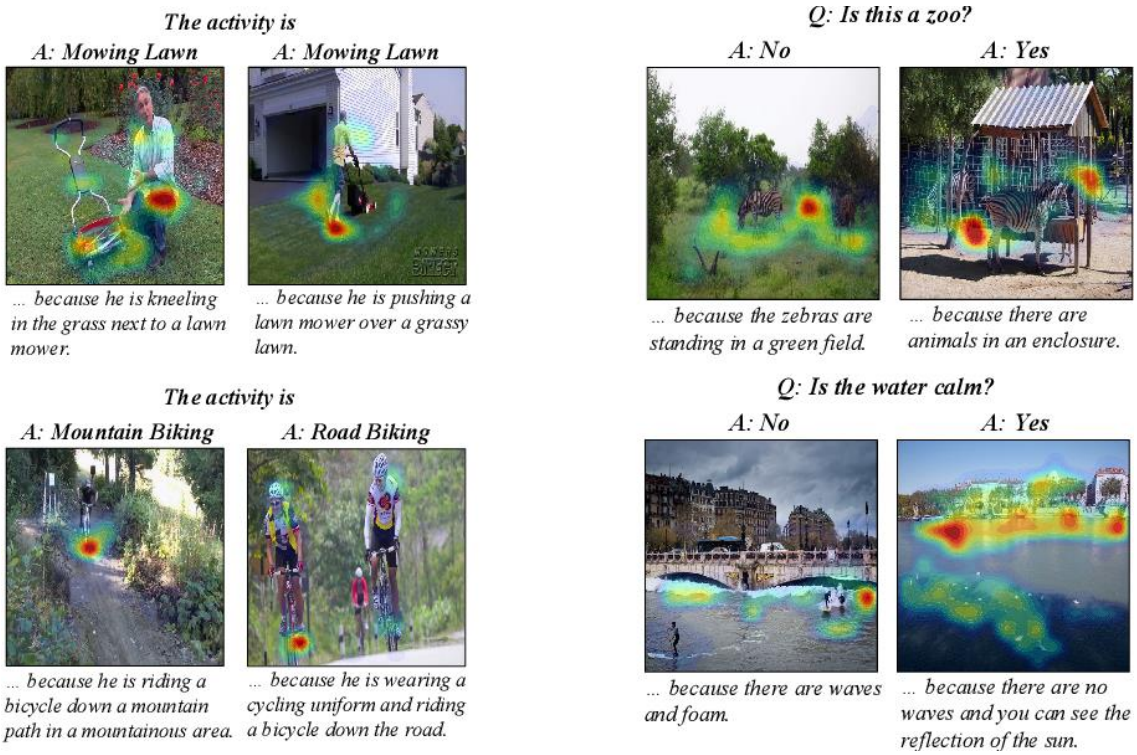
préciser quels critères sont utilisés pour garantir qu'il est équitable[226], et il faut être capable de le rendre indifférent à certains critères comme le genre ou la couleur de peau.

Ne pas connaître les critères identifiés par l'algorithme, c'est aussi ne pas savoir s'il attribue une classe *pour les bonnes raisons.* Les *data scientists* aiment illustrer ce point avec la légende de Hans le malin (« Clever Hans »), le cheval qui savait compter. Son dresseur énonçait une addition et il tapait du sabot jusqu'au nombre exact. Un comité d'experts démontra que l'animal ne comprenait pas les instructions mais percevait si bien les attentes du public qu'il pouvait en déduire quand arrêter de taper du sabot. Le cheval était capable de donner la bonne réponse, mais pour de mauvaises raisons[227]. De la même manière, un algorithme peut se calibrer sur les mauvaises caractéristiques. Par exemple, si tous les exemples de loups d'une base de données se tiennent sur la neige, un chien sur la neige risque d'être classé comme « loup » car l'algorithme donnera plus plus de poids à la présence de neige qu'aux traits distinctifs du loup. Sous le nom d'*explainable AI,* tout un pan de la recherche actuelle en intelligence artificielle vise à inventer et raffiner des méthodes permettant de faire apparaître les « raisons » amenant à un résultat, par exemple sous la forme de « cartes de chaleurs » sur une image montrant quelles zones ont le plus compté dans la classification opérée par le réseau de neurones (la neige ou bien la forme de l'animal), assorti d'une explication en langage naturel[228].

226 C'est ce qui a été exigé, par exemple, des algorithmes de *credit scoring* de Goldman Sachs, après qu'ils aient accusés de discriminer les femmes. Ian Carlos Campbelle, « The Apple Card doesn't actually discriminate against women, investigators say », *The Verge*, 23 mars 2021, https://www.theverge.com/2021/3/23/22347127/goldman-sachs-apple-card-no-gender-discrimination page consultée le 20 septembre 2021.

227 L'épisode est fréquemment mentionné. À titre d'exemple : Lapuschkin, Wäldchen, Binder, *et al.*, « Unmasking Clever Hans predictors and assessing what machines really learn », *Nature Communications,* 10, 1096, 2019.

228 Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, Andreas Holzinger, « Explainable AI : The New 42? », *CD-Make,* 2018.

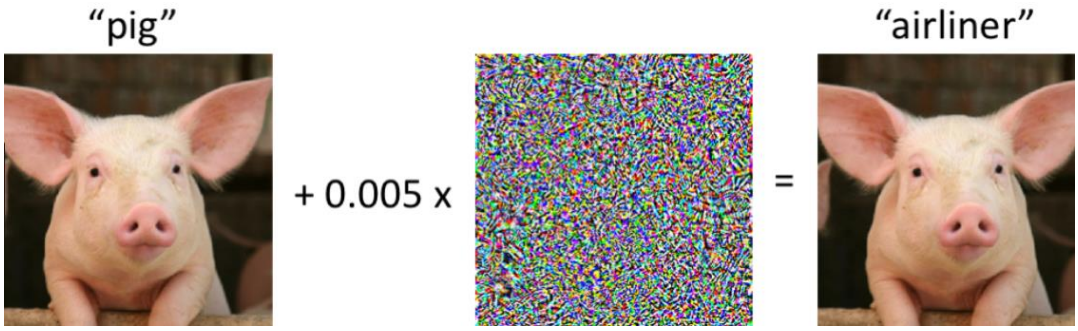


L'outil d'explication indique par une coloration de l'image les pixels qui ont été déterminants et fournit une explication en langage naturel. Source : Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, Andreas Holzinger, « Explainable AI : The New 42? », CD-Make, 2018.

Parmi les autres méthodes disponibles, la « feature visualisation » donne à voir, pour chaque couche du réseau, quels sont les critères déterminants, tandis que les « ablations studies » suppriment certaines parties du réseau pour évaluer leur contribution. Cependant, aucune méthode ne permet d'explication complète[229]. Pour certains chercheurs, il vaut mieux renoncer à une explication (exposer la procédure qui amène à un résultat) et lui préférer une justification (fournir des éléments montrant pourquoi le résultat est crédible)[230] – en combinant les méthodes disponibles[231].

Une des manières d'obtenir les cartes de chaleur est de modifier un à un les pixels de l'image et la soumettant à chaque fois au réseau de neurone pour déceler quel changement de pixel entraîne un changement de classification. En élaborant ces méthodes, les chercheurs ont eu la surprise de constater qu'il suffisait de changer certains pixels de ces zones décisives pour obtenir une classe complètement différente – alors que pour un œil humain, l'image est identique.

229 Olah, et al., « Feature Visualization », *Distill*, 2017.
230 Or Biran, Courtenay V. Cotton, « Explanation and Justification in Machine Learning : A Survey », 2017.
231 Olah, et al., « The Building Blocks of Interpretability », *Distill*, 2018.



La modification de quelques pixels de cette image de cochon amène le réseau de neurones à la classer comme « airliner » alors que pour l'oeil humain il s'agit toujours d'un cochon. Image : Kirthi Shankar Sivamani, « The Unusual effectiveness of adversarial attacks », Medium, *31 juillet 2019, https://medium.com/@smkirthishankar/the-unusual-effectiveness-of-adversarial-attacks-e1314d0fa4d3, page consultée le 20 septembre 2020.*

L'opacité constitue-t-elle un obstacle à l'ambition du projet d'intelligence artificielle ? Peut-on dire qu'il y a élucidation des mécanismes de l'intelligence si les machines inventées sont des « boîtes noires » ? Il est possible d'échapper à cette critique en comparant le *deep learning* avec l'intuition : la réponse peut être bonne (ou fausse), sans que l'on sache pourquoi. Nous n'avons pas accès à la complexité de l'interdépendance des paramètres du réseau de neurones, pas plus qu'aux « mécanismes inconscients » de notre propre cerveau. Avec le *deep learning,* comme avec l'intuition, nous gagnons en puissance ce que nous perdons en explicabilité et en vulnérabilité des résultats. Reste à savoir, si l'explication de la « boîte noire » de l'intuition est une autre « boîte noire », par quel moyen évaluer sa pertinence. Nous y reviendrons ultérieurement.

1.3.4. Des machines singulières ? « Magie noire » et problèmes de reproductibilité

Pour éviter toute confusion avec la section précédente, précisons à nouveau que les « paramètres » sont obtenus par le programme à la suite de sa calibration sur la base de données d'exemples, tandis que les « hyperparamètres » sont ceux que définit le programmeur avant l'apprentissage (par exemple, le nombre de couches du réseau de neurones, la largeur des couches, le type de fonction d'activation, le *learning rate*…). En plus des difficultés d'explicabilité des paramètres obtenus, que nous venons d'exposer, il existe une deuxième zone d'opacité au niveau des hyperparamètres fixés par le programmeur.

Nous avons vu que l'usage des algorithmes d'apprentissage (et en particulier des réseaux de neurones les plus récents) implique une certaine dose de tâtonnements. Les praticiens essayent différents hyperparamètres, voire différents algorithmes, et retiennent les plus efficaces. Avec l'expérience, ils acquièrent un ensemble de bonnes pratiques pour lequel ils n'ont pas de justification théorique : ils constatent que cela fonctionne mieux, mais sans savoir pourquoi. C'est la « magie noire » du *machine learning,* et en particulier du *deep learning.* Certains travaux visent à confier ce tâtonnement manuel à un autre algorithme qui compare les résultats des différents hyperparamètres et retient les meilleurs[232]. Mais dans la majorité des cas, l'obtention du meilleur résultat impose une partie d'ajustement manuel[233].

À cela s'ajoute le rôle joué par la base de données d'entraînement. Les algorithmes d'apprentissage se *singularisent* en se calibrant sur la base de données d'entraînement, et en récupèrent ainsi les « biais ». Nous l'avons mentionné, un algorithme de détection des mélanomes calibré sur une majorité de peaux blanches ne sera pas adapté aux peaux foncées. L'actualité ne cesse de rapporter les cas où des programmes calibrés sur des humains standards (homme, blanc, etc.) en viennent à dysfonctionner (et donc à discriminer) celles et ceux qui s'en écartent : algorithmes de reconnaissance faciale, d'attribution de crédit, de calcul de franchise d'assurance[234]…

Ainsi, le *machine learning* est loin d'offrir des algorithmes capables d'« apprendre » dans n'importe quel contexte. Avec l'**entraînement** (en particulier le choix et l'élaboration de la base de données) et le **paramétrage**, il y a un travail d'**ajustement** « manuel » – au sens où il est effectué par des humains qui, pour ce faire, doivent avoir une idée précise de la tâche à accomplir, du « sens » du problème à résoudre.

Cette difficulté se traduit également par une « crise de reproductibilité[235] » dans le champ du *machine learning.* Lors de la publication d'articles, il est difficile de rendre explicite l'ensemble des paramètres amenant au résultat : en sus du code, il faudrait mettre à disposition la base de données d'entraînement, ainsi que l'ensemble des hyperparamètres. Et encore, cela ne suffirait pas à garantir l'exactitude de la reproduction puisque certains programmes utilisent un état aléatoire comme point de départ de la calibration. Même en reproduisant à l'identique

232 Le champ a été baptisé « AutoML » pour « Automatic Machine Learning ». Voir par exemple Mendoza, Klein, Feurer, Springenberg, Hutter, « Towards Automatically-Tuned Neural Networks », *JMLR: Workshop and Conference Proceedings* 64:58–65, 2016.

233 Idriss Brahimi, « Un premier retour d'expérience sur AutoML », *Ysance Blog,* 28 septembre 2020, https://blog.ysance.com/un-premier-retour-experience-sur-automl, page consultée le 20 septembre 2021.

234 Kate Crawford et son équipe du *AI Now Institute* les recensent dans leurs rapports. Voir https://ainowinstitute.org/reports.html, page consulté le 1er septembre 2020.

235 Matthew Huston, « Artificial intelligence faces reproducibility crisis », *Science,* 16 février 2018, Vol 358, Issue 6377, p. 725-726.

les paramètres et la base de données, chaque entraînement donne des résultats légèrement différents.

L'usage du code laisse penser que nous avons affaire à une méthode **universelle**. Il semble plutôt que la **singularité** des problèmes impose un ajustement toujours différent, qu'il faut renouveler en cas de « dérive des concepts », et plus généralement face à un changement de contexte. Ainsi, lors des confinements de mars 2020, le bouleversement des comportements a mis en échec les algorithmes qui accompagnent la vente en ligne (recommandation, prévision des ventes, prévision des stocks…)[236]. Les algorithmes n'ont pas su « apprendre » de la situation comme nous l'avons fait. Au-delà d'une « crise de reproductibilité », les algorithmes se heurtent à l'impossible reproductibilité des crises – nous y revenons dans la troisième partie.

1.3.5. L'apprentissage profond, un retour aux sources ?

Selon Geoffrey Hinton, le *deep learning* est un retour aux sources du projet d'intelligence artificielle :

> Au tout début, dans les années cinquante, des gens comme von Neumann ou Turing ne croyaient pas dans l'IA symbolique, ils étaient bien plus inspirés par le cerveau. Malheureusement, ils sont tous les deux morts trop jeunes. Et leurs voix n'ont pas été entendues. Dans les premiers temps de l'IA, les gens étaient entièrement convaincus que les représentations requises par l'intelligence étaient des sortes d'expressions symboliques. Une sorte de logique […]. Et que l'essence de l'intelligence était le raisonnement[237].

Geoffrey Hinton positionne le *deep learning* comme le seul descendant légitime des pères fondateurs du projet d'intelligence artificielle. D'après lui, l'école symbolique n'a pu prospérer

236 Will Douglas Heaven, « Our weird behavior during the pandemic is messing with AI models », *MIT Technology Review,* 11 mai 2020, https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/? Page consultée le 1er septembre 2020.

237 « In the early days, back in the fifties, people like von Neumann and Turing didn't believe in symbolic AI, they were far more inspired by the brain. Unfortunately, they both died much too young. And their voice wasn't heard. And in the early day of AI people were completely convinced that the representations you needed for intelligence were symbolic expressions of some kind. Sort of cleaned up logic. Where you could do non-monotonic things and… Not quite logic, but something like logic. And that the essence of intelligence was reasoning. » Geoffrey Hinton, interviewé par Andrew Ng, « Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton », https://www.youtube.com/watch?v=-eyhCTvrEtE, page consultée le 4 octobre 2019.

qu'à la faveur de leurs décès précoces. Pourtant, les références à Turing et von Neumann sont discutables.

S'il est vrai qu'après la guerre Turing confiait son ambition de « construire un cerveau[238] » et qu'il aimait à désigner les premiers ordinateurs comme des « cerveaux électroniques[239] », sa notion de « cerveau » avait plus à voir avec les « états d'esprits » évoqués dans l'article de 1936 qu'avec la réalité physiologique du cerveau[240]. Il s'agissait de construire une machine universelle en ignorant les détails de la physiologie pour privilégier la fertilité de la notion de programme, « ce qui mettait l'accent non sur le fonctionnement interne du cerveau, mais sur les instructions explicites qu'un travailleur humain pouvait suivre à la lettre »[241]. Il faut donc faire l'impasse sur une partie de son œuvre pour présenter Turing comme un précurseur du connexionnisme – partie qu'il suffirait d'exhumer pour permettre à l'école symbolique de défendre le point de vue opposé. Il est plus pertinent de positionner Turing *en amont* de la distinction entre écoles. Nous avons mentionné la fin de l'article de 1950, où Turing s'interroge : faut-il commencer par apprendre aux machines à jouer aux échecs, qui est une voie plus « rationaliste » ou « symbolique », ou bien suivre une voie plus « empiriste » en les dotant d'organes sensoriels ? On se souvient de sa conclusion : « je pense qu'il faudrait essayer les deux voies[242] ».

La référence à von Neumann est tout aussi fragile. Son intérêt pour la comparaison entre l'ordinateur et le cerveau peut le faire passer pour un connexionniste avant l'heure. Pour autant, il s'est montré critique vis-à-vis de ce que les chercheurs en informatique ont retenu du neurone. Considérer les neurones comme des portes logiques est simpliste, et laisse de côté d'autres aspects importants des neurones réels[243].

Ainsi, en choisissant l'agencement entre neurones comme niveau de description, les connexionnistes se distinguent à la fois de Turing, pour qui le bon niveau est « plus haut », à l'échelle des processus de pensée (les « états d'esprits »), et de von Neumann, pour qui il ne faudrait pas ignorer ce qui se passe « plus bas », dans le détail du fonctionnement des neurones.

238 Andrew Hodges, *Alan Turing ou l'énigme de l'intelligence*, *op. cit.*, p. 247.
239 *Ibid*, p. 251.
240 « Pour notre mathématicien, quoi que fasse un cerveau, il le faisait en vertu de sa structuration logique et non parce qu'il se trouvait à l'intérieur d'un crâne humain ou parce qu'il était constitué de matière spongieuse composée d'une espèce particulière de formation cellulaire biologique. Sa structure logique devait parfaitement être réplicable dans un autre milieu, matérialisée par une autre espèce de mécanisme physique. » *Ibid*, p. 248.
241 *Ibid*, p. 248.
242 Alan Turing, « Les ordinateurs et l'intelligence », *op. cit.*, p. 175.
243 John von Neumann, *L'ordinateur et le cerveau, op. cit*.

1.3.6. « It will come to the same thing in the end »

En revendiquant l'héritage intellectuel de Turing et von Neumann, Hinton a pour but principal de discréditer l'école symbolique. Mais la revendication ne tient pas : il est facile, pour un tenant de l'école symbolique, de trouver chez Turing ou von Neumann de quoi prouver la même filiation. C'est que, si les écoles symboliques et connexionnistes s'opposent quant au niveau de description, elles partagent la même ambition et la même hypothèse fondamentale, nées à la génération de Turing et von Neumann et formulées le plus clairement à la génération suivante, à l'occasion du séminaire de Dartmouth : **il existe un bon niveau de description, et celui-ci permettra d'élucider et reproduire l'intelligence**.

Ainsi, il ne fait aucun doute pour Geoffrey Hinton que son travail permettra à terme d'élucider les mécanismes de l'esprit. Pour lui, « une pensée, c'est juste un sacré gros vecteur d'activité neurale[244] ». Il y a équivalence entre l'état du cerveau et la pensée, la situation de chaque neurone correspondant aux paramètres de la pensée en cours. Il se moque de l'école symbolique, pour qui :

> ce qui rentre est une suite de mots, et ce qui sort est une suite de mots, et à cause de ça les suites de mots sont la façon la plus évidente de représenter les choses, donc ils pensaient que ce qu'il y a entre les deux devait être une suite de mots, ou quelque chose comme une suite de mots[245],

une conception qui est pour lui « la plus bête des idées ». Geoffrey Hinton préfère remplacer les « suites de mots » par « de gros vecteurs », ce qui, précise-t-il, « est totalement différent de la vue standard de l'IA selon laquelle les pensées sont des expressions symboliques[246] ».

Hinton met l'accent sur son opposition à l'école symbolique, mais il partage la même ambition d'expliquer l'intelligence, que ce soit par une « suite de mots » ou par de « gros vecteurs ». La divergence entre école connexionniste et école symbolique se situe au niveau de

244 Nous traduisons : « what a thought is, is just a great big vector of neural activity. » Geoffrey Hinton, interviewé par Andrew Ng, « Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton », *op. cit.*

245 Le texte original est : « what comes in is a string of words, and what comes out is a string of words, and because of that strings of words are most obvious way to represent things, so they thought what must be in between was a string of words, or something like a string of words, and I think what is in between is nothing like a string of words. I think the idea that thought must be in some kind of language is a silliest idea that understaning the layer in a special scene must be in pixels. » *Ibid.*

246 « thoughts are just big vectores, and the big vectors have causal powers, they cause other big vectors, and that's utterly unlike the standard AI view that thoughts are symbolic expressions. » *Ibid.*

la méthode mais pas au niveau du but. Conformément à la conjecture formulée à Dartmouth, l'objectif reste de dissiper les « mystères » de l'intelligence.

Il est courant de présenter l'école connexionniste comme un dépassement de l'école symbolique, et plus spécifiquement de l'idée d'une pensée caractérisée par un ensemble de règles opérant sur des symboles. L'école connexionniste se targue de révéler les mécanismes sous-jacents dont les « raisons » ne seraient que le résultat émergent. Mais que l'on cherche le mécanisme de la pensée au niveau de l'enchaînement des symboles ou au niveau de l'interaction des neurones, il s'agit bien, dans les deux cas, de chercher un mécanisme. Ainsi que le formule Yann LeCun, l'objectif des chercheurs contemporains est de « construire une machine intelligente pour savoir quels sont les principes sous-jacents qui sont vraiment importants[247] ». Le but est de trouver le système de règles qui permette de rendre compte de l'intelligence. Pour Yann Lecun, cela se traduit par un modèle de l'esprit – en cours d'élaboration –, qui servira à la fois à fabriquer des machines intelligentes et à décrire notre propre intelligence. Le modèle doit intégrer un certain nombre de « modules » : mémoire, prévision de l'état du monde, circuit de récompense… et décrire comment ces modules interagissent les uns avec les autres.

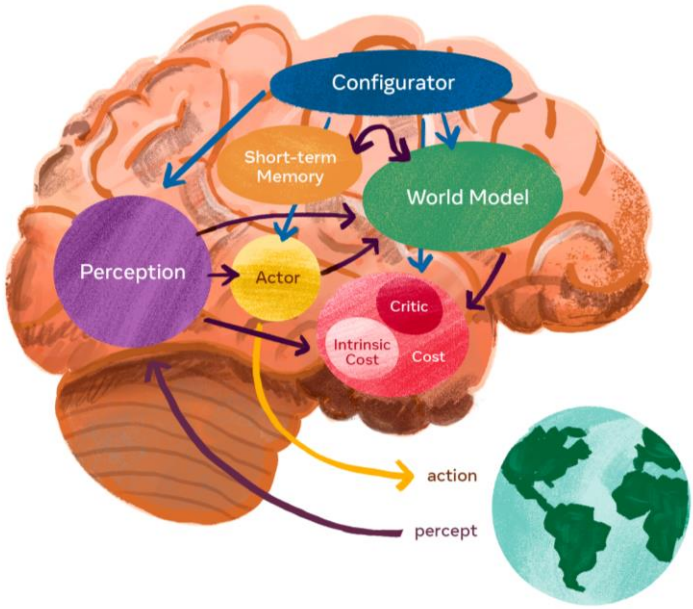


Image : modèle du raisonnement selon Yann LeCun. Source « Yann LeCun on a vision to make AI systems learn and reason like animals and humans », Meta AI Blog, *23 février 2022. https://ai.facebook.com/blog/yann-lecun-advances-in-ai-research/ page consultée le 22 avril 2022.*

247 Yann LeCun, « Les émotions sont inséparables de l'intelligence », interview sur France Culture, mis en ligne sur *YouTube* le 18 octobre 2018, https://www.youtube.com/watch?v=ZvMpTm_AXQM, page consultée le 4 octobre 2019.

Si elles s'opposent sur le niveau de description et la méthode, l'école symbolique comme l'école connexionniste partagent un certain nombre d'hypothèses : qu'il est possible d'élaborer un modèle de l'intelligence (sous la forme d'un système de calcul logique ou d'un agencement de neurones), que celui-ci repose sur des unités élémentaires (qu'il s'agisse d'une proposition logique ou d'un neurone), et consiste en du calcul opéré sur ces unités élémentaires. David Bates ne s'y trompe pas lorsqu'il évoque l'« hypothèse sous-jacente » qui « fonde à la fois la recherche traditionnelle en IA […] et les nouvelles formes d'investigation […] : Les deux approches croient que la pensée sera répliquée (en théorie, du moins), une fois que ces processus sous-jacents seront révélés[248] ». Or, cette hypothèse sous-jacente correspond précisément à celle qui a été formulée à l'occasion du séminaire de Dartmouth, par ceux qui sont devenus ensuite les principaux champions de l'école symbolique : tous les aspects de l'intelligence peuvent être décrits avec une précision telle qu'une machine peut les simuler[249]. Autrement dit, l'école connexionniste continue à porter haut le flambeau de la conjecture de Dartmouth – pourtant formulée par leurs ennemis de l'école symbolique.

Cet accord implicite avait été déjà été décelé par Walter Pitts, avant le séminaire de Dartmouth et avant que les deux écoles ne se constituent, à l'occasion d'un séminaire à la RAND en 1954. Newell présente un programme de jeux d'échecs tandis que Selfridge expose un programme de reconnaissance de formes. Dans l'assistance, Walter Pitts souligne la différence entre les deux approches : les uns « imitent le système nerveux » tandis que les autres « préfèrent imiter la hierarchie des causes finales traditionnellement appelées l'esprit ». Mais pour lui « cela reviendra au même au final, sans aucun doute[250]… ».

Il était aisé pour Walter Pitts de n'y voir que « différents niveaux de description » qui pourraient converger puisque le modèle qu'il avait proposé avec McCulloch consistait *à la fois* en un modèle des processus à l'œuvre dans le cerveau et des processus mentaux en tant que

248 « What I want to do here is question the underlying assumption that grounds both traditional AI research […] and newer forms of investigation into intelligence […]. Both approaches believe that thought will be replicated (theoretically at least) once these underlying processes are revealed. » David Bates, « Creating Insight : Gestalt Theory and the Early Computer » in Jessica Riskin (ed.), *Genesis Redux, Essays in the History and Philosophy of Artificial Life*, Chicago, The University of Chicago Press, 2007, p. 238-239.

249 Pour rappel : « The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. » McCarthy, J. Minsky M. L, Rochester, N., Shannon C.E. , « A proposal for the Dartmouth summer research project on artificial intelligence », *op. cit.*

250 « Walter Pitts, a commentator for this session, concluded it by saying, "But, whereas Messrs. Farley, Clark, Selfridge, and Dinneen are imitating the nervous system, Mr. Newell prefers to imitate the hierarchy of final causes traditionally called the mind. It will come to the same thing in the end, no doubt…' To "come to the same thing," these two approaches, neural modeling and symbol processing, must be recognized simply as different levels of description of what goes on in the brain. » Nils J. Nilsson, The Quest for Artificial Intelligence, A History of Ideas and Achievements, Cambridge, Cambridge University Press, 2010, p. 76-77.

succession d'idées reliées par des règles logiques. L'article de 1943 fait mention de « psychons » comme étant les unités psychiques de base, à la fois opérateurs de la logique des propositions et associées à l'activité des neurones[251]. Selon les termes d'aujourd'hui, leur modèle est donc *à la fois* connexionniste (il s'inspire du cerveau et des réseaux de neurones) et symbolique (l'état de chaque neurone correspond à une proposition logique et participe à un calcul propositionnel).

1.3.7. L'IA contre la cybernétique, tout contre

Les connexionnistes se trouvent dans une situation épistémologique délicate. Ils reprennent *à la fois* l'objectif historique du projet d'intelligence artificielle – étudier et réaliser l'intelligence –, et les méthodes de la cybernétique : ils fabriquent des machines apprenant par essai-erreur sur une base de données d'exemples, ou par *feedback* de l'environnement. Comme le remarque Michael Jordan, « sous la bannière de la terminologie de McCarthy » (réaliser l'intelligence) on retrouve « l'agenda intellectuel de Wiener » (la calibration par feedback)[252].

Or, le projet d'intelligence artificielle s'est justement constitué en réaction contre la cybernétique. Cette dernière reposait sur « une représentation physicaliste de l'information comme expression de la quantité d'ordre ou de structure dans les agencements matériels[253] » – autrement dit une théorie du *signal* –, ne permettant pas d'isoler un niveau du traitement de l'information indépendant des processus matériels. Pour les fondateurs du projet d'intelligence artificielle, une telle notion d'information empêche de s'atteler à une « théorie de l'esprit[254] ». Il faut donc séparer la matière de l'information, concevoir cette dernière comme *code* plutôt que comme *signal*[255]. Ainsi, le projet d'intelligence artificielle s'attache à « isoler un niveau

251 « To psychology, however defined, specification of the net would contribute all that could be achieved in that field even if the analysis were pushed to ultimate psychic units or "psychons," for a psychon can be no less than the activity of a single neuron. Since that activity is inherently propositional, all psychic events have an intentional, or "semiotic," character. The "all-or-none" law of these activities, and the conformity of their relations to those of the logic of propositions, insure that the relations of psychons are those of the two-valued logic of propositions. Thus in psychology, introspective, behavioristic or physiological, the fundamental relations are those of two-valued logic.» McCulloch et Pitts, *op. cit.,* p. 131

252 Michael Jordan, « Artificial Intelligence — The Revolution Hasn't Happened Yet », publié sur *Medium* le 19 avril 2018, https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7, page consultée le 4 octobre 2019. Nous traduisons.

253 Mathieu Triclot, *Le moment cybernétique : La constitution de la notion d'information,* Seyssel, Éditions Champ Vallon, 2008, p. 10.

254 « L'erreur de la cybernétique tient toute en un seul point pour Marvin Minsky : ne pas avoir fait une théorie de l'esprit », *Ibid*.

255 Shannon a été un participant actif du séminaire de Dartmouth, et un des quatre signataires de l'appel à projet.

autonome [c'est-à-dire *indifférent au support*], celui des représentations[256] » dans le but d'abstraire de la matière un hypothétique *logiciel de l'intelligenc*e.

Les connexionnistes se sont fait le relai de l'ambition d'arriver à une théorie de l'intelligence et d'en dissiper les « mystères ». Mais les moyens qu'ils se donnent sont-ils conformes à ce but ? En adoptant la méthode cybernétique d'une machine se calibrant par *feedback*, l'école connexionniste semble se priver de la perspective d'une théorie, puisqu'après entraînement, les chercheurs se trouvent face à une « boîte noire ». A partir du moment où les machines se paramètrent d'elles-mêmes à partir d'une base de données (apprentissage supervisé) ou par interaction avec l'environnement (apprentissage par renforcement), il devient difficile de les étudier pour en tirer une « théorie » de l'intelligence.

Lors des conférences Macy, von Neumann avait anticipé cette situation. « Bientôt, prophétis[ait]-il, le constructeur d'automate sera aussi désarmé devant sa création que nous le sommes devant les phénomènes complexes[257]. » Si l'objet que l'on cherche à simuler est trop complexe, la simulation que nous en faisons ne risque-t-elle pas d'être tout aussi complexe ? Jean-Pierre Dupuy commente la remarque de von Neumann :

> Le modèle, qui était hiérarchiquement subordonné au réel qu'il ne faisait que mimer, s'émancipe et devient l'égal de son référent. […] Mais alors, il sera non seulement modèle de son objet, mais aussi modèle de lui-même, ou plutôt de son comportement. Il n'en faut pas davantage pour que le modèle ne renvoie plus qu'à lui-même, polarisant sur lui l'attention, au détriment de l'objet originel[258].

Au lieu de nous révéler le fonctionnement de l'objet étudié, le modèle devient un objet d'étude à part entière. Il ne rend pas notre vision des choses plus simple, mais la complique. Jean-Pierre Dupuy conclut : « C'est ainsi que la neurophysiologie laisse la place à l'intelligence artificielle[259]. » Cela s'applique particulièrement à sa mouvance connexionniste : « Le néoconnexionniste se rapporte à son réseau comme à un objet complexe au sens de von Neumann, quasi naturel ; il en étudie expérimentalement les propriétés, le manipule, fait varier ses paramètres, etc[260]. »

Contre von Neumann et sa conception pragmatique de la notion de modèle, McCulloch défend une conception qui préfigure l'obstination théorique des connexionnistes

256 Mathieu Triclot, *Ibid.*
257 Jean-Pierre Dupuy, *op. cit.*, p. 156.
258 *Ibid.*
259 *Ibid.*
260 *Ibid*, p. 167.

contemporaires. Son réseau de neurones éclaire le fonctionnement réel du cerveau tout en étant plus simple. Ce sont deux façons divergentes de penser la notion de modèle : pour McCulloch, un modèle **abstrait**, il distingue le nécessaire de l'accidentel et néglige les détails pour ne garder que ce qui permet de fonctionner. Pour von Neumann, un modèle **simule**, et prend en compte les détails, quitte à devenir aussi imprévisible que ce qu'il simule.

Si les inventeurs du *deep learning* se réclament de l'hypothèse de Dartmouth et pensent rendre explicites les mécanismes de l'intelligence, leurs réalisations s'écartent de cette ambition. Les algorithmes créés sont d'une efficacité sans précédent pour effectuer les tâches qui leur sont affectées, mais leur complexité ne semble pas permettre de satisfaire l'impératif d'élucidation de l'intelligence. Ils fabriquent des simulations, tout en leur revendiquant le statut d'abstraction.

1.3.8. De la pratique à la théorie, le principe du *verum factum*

Tout se passe comme s'il fallait choisir entre d'une part, l'abstraction et l'intelligibilité, et d'autre part, la concrétisation et l'opacité. Il faudrait négliger la matière pour une *modélisation* claire, ou bien renoncer à la compréhension pour une *opération* efficace. Dès lors, il serait paradoxal de se réclamer *à la fois* de l'une et de l'autre, de la conjecture de Dartmouth et d'un retour à la cybernétique, d'une quête théorique du logiciel de l'intelligence et de la mise au point de machines intelligentes.

Ce n'est pourtant pas le point de vue des chercheurs. Un de leur *leitmotiv,* est d'affirmer que le meilleur moyen de *savoir* est de *faire,* et qu'il existe de nombreux exemples où la pratique a précédé la théorie. On se souvient des exemples mentionnés par Yann LeCun : la machine à vapeur a été inventée bien avant la thermodynamie, et les premiers avions ont volé avant que ne soit théorisée la portance (voir supra 1.3.11.).

L'optimisme des chercheurs se fonde sur une adhésion sans réserve à une lecture contemporaine du principe déjà mentionné du *verum factum* : « *Verum et factum convertuntur* » (« Ce qui est vrai et ce que l'on fait sont convertibles »). Dans sa première lecture, le principe vise à montrer que Dieu connaît parfaitement la nature, puisqu'il l'a créée, contrairement à

l'humain, qui ne peut que l'observer[261]. Puis, le principe a pris un sens plus centré sur l'humain : « Ce que l'homme fait, il peut le connaître rationnellement, de façon démonstrative et déductive, malgré la finitude de son entendement[262]. » Aujourd'hui, il est devenu une règle de recherche : « on ne peut connaître qu'en faisant, ou plutôt qu'en re-faisant[263] ». C'est cette dernière formulation que Yann LeCun répète à l'envi, en l'attribuant à Richard Feynman[264] . De même, la fabrication de machines intelligentes pourrait donner lieu à une théorie de l'intelligence. La théorie aura peut-être *un temps de retard* sur la pratique, mais elle finira par arriver. La *théorie* est donc bien l'horizon du projet de *fabrication*.

1.3.9. Récapitulatif

Pour clarifier ces paradoxes, nous en récapitulons les points principaux :

1 Les connexionnistes contemporains se présentent comme les ennemis de l'école symbolique mais ils revendiquent la même chose : expliquer et répliquer l'intelligence.

2 En cela, ils sont fidèles à la conjecture formulée à l'occasion de Dartmouth par ceux qui sont devenus ensuite les principaux champions de l'école symbolique : tous les aspects de l'intelligence peuvent être décrits avec une précision telle qu'une machine peut les simuler.

3 Mais les machines fabriquées par l'école connexionniste ne satisfont pas l'ambition d'élucidation de l'intelligence : elles sont mises au point par tâtonnements et se calibrent sur la singularité des situations. Elles se rapprochent des caractéristiques du *signal* (singularité, matérialité, opacité), plutôt que de celles du *code* (universalité,

261 Énoncé par Vico, le principe lui permet de comparer les savoirs relatifs à la physique (fabriquée par Dieu) de ceux relatifs aux affaires humaines (fabriquées par les humains). C'est aussi une manière d'affirmer que la « nature » d'une chose n'est pas une « essence cachée » mais seulement les circonstances de sa production.

262 Jean-Pierre Dupuy, *Aux origines des sciences cognitives*, Paris, La Découverte, 2005, p. 17.

263 *Ibid.*

264 « le principe [est] érigé par le physicien célèbre Richard Feynman qui a dit 'on ne comprend pas vraiment quelque chose tant qu'on ne l'a pas construit soi-même' », Yann LeCun, « Les émotions sont inséparables de l'intelligence », *YouTube,* 18 octobre 2019, minute 4. Cade Metz rapporte également cette influence de Feynman auprès de Bengio et Hinton. « The blackboard in Feyman's classroom once read: "What I cannot create, I do not understand." This was what Yoshua Bengio, Goodfellow's advisor at the University of Montreal, had argued in a café near the university when wooed by the traveling contingent from Microsoft. Like Hinton, Bengio and Goodfellow believed Feyman's adage applied to machines as well as people: What artificial intelligence cannot create, it cannot understand. » Cade Metz, *Genius Makers, The Mavericks Who Brought AI to Google, Facebook, and the World,* New-York, Penguin Random House, 2021, chapitre 13.

indépendance par rapport au support, transparence). Elles n'offrent pas la théorie promise.

4 En cela, l'école connexionniste effectue un retour aux sources, à celles de la cybernétique plutôt qu'à celles du projet d'intelligence artificielle. Par conséquent, l'école connexionniste est prise dans une contradiction entre la volonté affichée d'aboutir à une « théorie de l'esprit » et les artefacts réalisés, qui défient la compréhension.

5 Pour concilier leurs machines singulières et la volonté d'élucidation, l'école connexionniste en appelle au principe du *verum factum* : fabriquer est une manière de comprendre. Pour l'instant, les chercheurs ont réussi à mettre au point les machines, l'explication suivra dans la foulée.

On pourrait attendre des chercheurs que, conformément aux prétentions affichées, ils travaillent à *interpréter* la pertinence de leurs outils en tant que modèles de l'intelligence. Une fois qu'une machine fonctionne, il faudrait se demander ce qu'elle permet de comprendre de l'intelligence. Or, à l'exception de quelques déclarations à l'emporte-pièce, l'essentiel de leurs efforts va vers *l'amélioration* des inventions, plutôt que vers leur *interprétation.* Qu'il s'agisse de Geoffrey Hinton ou de Yann LeCun, le succès de leurs inventions ne les a pas conduits à une entreprise d'élucidation, mais à s'atteler à de nouvelles inventions : aujourd'hui, Geoffrey Hinton s'intéresse aux *capsules networks* et Yann LeCun à l'apprentissage auto-supervisé. Tout comme, à la fin de l'article de 1950, Turing ne répond pas à la question de savoir si les machines peuvent penser, les chercheurs ne répondent pas à leur question initiale (« nos machines pensent-elles ? »), mais ne cessent de lancer de nouvelles pistes de recherches. Ce sont d'autres chercheurs qui enquêtent, à leur place, sur *ce que peut apporter le succès des réseaux de neurones à la compréhension de l'intelligence* – autrement dit, ce qu'ils valent en tant que « théorie de l'esprit ».

1.4. Décrire la pensée à l'aide de l'apprentissage profond

1.4.1. Les réseaux de neurones et le cerveau

Entre le dix-septième et le vingtième siècle, l'étude du cerveau s'est plutôt intéressée à sa plasticité, à sa capacité à se réorganiser en cas de choc, qu'aux éventuels traits communs qu'il pourrait avoir avec une machine déterminée[265]. En expliquant comment un ensemble de portes logiques organisées en réseau peuvent réaliser des opérations logiques, McCulloch et Pitts entreprennent, au contraire, de décrire le cerveau comme une machine[266]. Les critiques ont eu beau souligner l'extrême simplification de leur notion de neurone, cet argument n'a diminué en rien leur conviction. Pour modéliser un cerveau, il faut s'abstraire des détails et ne retenir que ce qui le fait fonctionner. Ainsi, leurs successeurs ont pu croire posséder la clef qui leur permettrait d'organiser d'autres éléments matériels *comme un cerveau* de manière à lui faire faire les mêmes opérations.

Comme exposé précédemment, l'invention des réseaux de neurones convolutionnels s'inspire à nouveau de la physiologie du cerveau en reprenant certains éléments des travaux de Hubel et Wiesel – le traitement hiérarchique du signal[267], ainsi que la distinction entre cellule simple et cellule complexe. Trente ans après, la distinction entre cellule simple et cellule complexe a été relativisée, montrant qu'une observation fausse ou inexacte en neurophysiologie pouvait constituer une inspiration féconde en intelligence artificielle. Malgré cela, la comparaison entre cerveau et réseau de neurones reste pertinente – en particulier en ce qui concerne la structure « hiérarchique » de la détection[268] et l'encodage des différences sous la

265 David Bates, « Automaticity, Plasticity, and the Deviant Origins of Artificial Intelligence », in Bates et Bassiri (eds.), *Plasticity and Pathology : On the Formation of the Neural Subject*, New York, Fordham University Press, 2015. Il mentionne l'observation par Descartes du cerveau comme un matière « molle et pliante » ainsi que les travaux de Lashley, von Monakow, Goldstein, James, de la *Gestalt theory* et jusqu'à ceux, contemporains de McCulloch et Pitts, de Ross Ashby, qui s'intéresse à une machine qui aurait, comme le cerveau, la possibilité de se réorganiser lorsqu'elle est mise en échec.

266 Jean-Pierre Dupuy, *op. cit.*, p. 53-54.

267 Pour rappel : les premières couches détectent certaines caractéristiques élémentaires (contraste, dégradé) et envoient les informations à la couche suivante où les neurones compilent ce qui a été détecté pour en déduire la présence de caractéristiques plus élaborées (lignes, frontières…). Le même processus est répété de couche en couche, permettant de détecter des objets à la complexité croissante.

268 Hong, Yamins, Majaj, *et al.* « Explicit information for category-orthogonal object properties increases along the ventral stream ». *Nature Neuroscience*, 19, 2016, p. 613–622.

forme d'une répartition dans un espace vectoriel[269]. Le succès du *deep learning* a conduit les neuroscientifiques à se demander si les progrès informatiques, bien qu'effectués indépendamment de la recherche en physiologie, peuvent éclairer la manière dont le cerveau fonctionne[270]. Les différences entre le cerveau et les réseaux de neurones ne sont pas négligeables : von Neumann mentionnait l'importance du détail des neurones, il faudrait ajouter aujourd'hui celui des axones, des neurotransmetteurs, de la glie… Il est difficile de tenir la position, comme McCulloch, selon laquelle les réseaux de neurones conservent l'essentiel du fonctionnement du cerveau, le reste pouvant être négligé. À cela s'ajoute les nombreuses différences de fonctionnement : le cerveau et les réseaux de neurones convolutionnels ne traitent pas les mêmes données et n'opèrent pas de la même façon : nous combinons les sens, nous utilisons l'attention, la mémoire[271], etc. Dès le symposium Hixon en 1948, Lashley adressait une objection de taille à McCulloch : le cerveau manifeste une activité spontanée permanente. Aussi, la perception n'est pas une activation des neurones par un *input* se traduisant par un *output* mais la modulation d'une activité préexistante[272]. Selon les mots de Jean-Pierre Changeux, elle peut être comparée à une horloge dont les aiguilles tournent déjà. Les stimulis viennent seulement « les retard[er] ou les remett[re] à l'heure[273] ». Cela explique pourquoi la *back-propagation*, procédé crucial de calibration des algorithmes d'apprentissage profond, ne semble pas s'appliquer au cerveau : il faudrait que notre activité cérébrale s'arrête et s'inverse, le temps d'ajuster le poids de nos neurones[274]. Les neuroscientifiques cherchent donc d'autres procédés, à même d'expliquer la calibration de nos neurones[275]. Enfin, les réseaux de neurones échouent à rendre compte de l'extraordinaire plasticité du cerveau. En cas de choc ou de traumatisme, le cerveau est capable de changer radicalement d'organisation, ce qui laisse supposer qu'il reste toujours en partie partie désorganisé. Pour William James, les parties les

269 Nikolaus Kriegeskorte, « Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing », *Annual Review of Vision Science* 1:1, 2015, p. 417-446.

270 Daniel Yamins, James J. DiCarlo. « Using goal-driven deep learning models to understand sensory cortex. » *Nature Neuroscience* 19, 2016, p. 356-365. Olivia Guest, Bradley C. Love, « Levels of Representation in a Deep Learning Model of Categorization », bioRxiv 626374, 2019.

271 Robert Jacobs et Christopher Bates. « Comparing the Visual Representations and Performance of Humans and Deep Neural Networks. » *Current Directions in Psychological Science,* 28, 2019, p. 34 - 39.

272 Jean-Pierre Dupuy, *op. cit.,* p. 146.

273 « Les organes des sens se comportent comme des 'commutateurs' d'horloges moléculaires. Les stimuli physiques qu'ils reçoivent du monde extérieur les avances, les retardent ou les remettent à l'heure. » Jean-Pierre Changeux, *L'homme neuronal,* Paris, Fayard, 1983, p. 107.

274 Lillicrap, Santoro, Marris, Akerman, Hinton, « Backpropagation and the brain », *Nature Reviews Neuroscience,* 21, 2020, p. 335–346.

275 Par exemple, Payeur, Guerguiev, Zenke *et al.* « Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits », *Nature Neuroscience,* 24, 2021, p. 1010–1019.

plus évoluées du cerveau étaient probablement les moins déterminées – leur imprécision les rendant disponible aux réorganisations[276].

Il est donc exagéré de voir le *deep learning* comme un modèle du fonctionnement du cerveau. Il y a, au mieux, une conversation entre neurophysiologie et intelligence artificielle, chaque discipline fournissant matière à inspiration pour l'autre. La conversation est toutefois assez discrète : les chercheurs en intelligence artificielle semblent déjà assez occupés par l'actualité trop foisonnante de leur propre champ pour suivre celle de la neurophysiologie[277]. Quant aux chercheurs en neurosciences, ils semblent plus intéressés par le *deep learning* en tant qu'instrument d'investigation permettant d'améliorer l'imagerie médicale et de tirer parti des données collectées[278], qu'en tant que modèle du fonctionnement du cerveau.

Cependant, les performances obtenues conduisent à admettre que les réseaux convolutionnels profonds méritent le qualificatif de « vision artificielle ». Bien qu'il soit aisé de souligner la différence entre ce qu'ils effectuent et l'expérience humaine ou animale de la vision, ils sont au moins capables de distinguer un objet d'un autre et de signaler leur présence dans une variété de contextes et de positions. Cette capacité n'est pas anodine, elle implique l'encodage par le réseau d'une représentation de l'objet robuste aux variations qui peut prétendre au statut de « concept ». De plus, l'acquisition de cette « représentation interne » à partir d'une collection d'exemples rapproche l'apprentissage de la notion d'« induction ». C'est ce que nous allons examiner à présent.

1.4.2. À quel type de jugement comparer le *deep learning* ?

Tout en se gardant d'attribuer des facultés psychologiques aux ordinateurs, Arno Schubbach pose la question suivante : si les algorithmes de *deep learning* sont des machines qui « jugent », de quel type de jugement font-elles preuve[279] ? Selon la *Critique de la raison pure* de Kant, les objets de la connaissance ne sont pas indépendants du processus de connaissance. Ils sont tributaires des catégories, c'est-à-dire d'un ensemble de règles auxquelles ils doivent se

276 David Bates, « Insight in the Age of Automation », *in* Joyce Chaplin, Darrin McMahon (dir.), *Genealogies of Genius*, Londres, Palgrave Macmillan, 2016, p. 162.

277 À l'exception notable de Hinton, co-signataire d'un article comparant l'apprentissage par *back-propagation* et le fonctionnement du cerveau.

278 À titre d'exemple, dans le diagnostic précoce de la maladie d'Alzheimer. Venugopalan, Tong, Hassanzadeh, *et al*. « Multimodal deep learning models for early detection of Alzheimer's disease stage » *Scientific Reports*, 11, 3254, 2021.

279 Arno Schubbach, « Judging machines: philosophical aspects of deep learning », *Synthese*, 2019, p. 1–21.

conformer pour pouvoir apparaître. Un jugement s'opère par synthèse des intuitions selon un concept conforme aux catégories. Pour Schubbach, les algorithmes classiques peuvent être comparés à cette définition du jugement. Les règles de traitement de l'information y sont définies par avance, tout comme les catégories sont fixées en amont des jugements (*a priori*). Les données sur lesquelles ils s'appliquent doivent se conformer à ces règles et ne les modifient pas.

Dans la *Critique de la faculté de juger,* Kant évoque certains domaines de la connaissance – comme la physique, la chimie ou la biologie – où les règles de traitement de l'objet ne préexistent pas. Il faut les découvrir par l'expérience, puis les valider en les éprouvant sur des objets de la même classe. De la même manière, remarque Schubbach, les algorithmes de *deep learning* adaptent leurs paramètres en fonction des relations qu'ils décèlent entre les objets de la base de données d'exemples, puis ces paramètres sont mis à l'épreuve par la base de données de test. Leur fonctionnement est déterminé par l'adaptation de leur structure (poids et biais) à la base de données.

Une fois induites des données, les règles découvertes ne deviennent pas explicites. La matrice des paramètres du réseau ne constitue pas une « explication » satisfaisante. Pour Schubbach, cela rapproche le *deep learning* de l'avis d'un expert humain auquel on fait confiance, sans forcément connaître les tenants et aboutissants de son raisonnement. Pour le *deep learning* comme pour l'expert, les raisons amenant à un avis, et son niveau de pertinence, demeurent inconnus. La confiance dans le jugement de l'expert se fonde sur son expérience et les circonstances de son jugement. La confiance dans le jugement du réseau de neurones se fonde sur la qualité des bases de données et de test, et sur les performances atteintes pendant l'entraînement. Dans les deux cas, il est possible d'évaluer l'entité qui juge, mais pas les avis singuliers. Un corpus de règles implicites acquises au fil de l'expérience produit des avis qui sont bons la plupart du temps, sans qu'il soit possible de l'évaluer précisément. Autrement dit, il est possible de *justifier*, par l'expérience, ou par d'autres méthodes, l'avis d'un expert, mais non de *l'expliquer,* d'en donner les raisons précises[280].

En guise d'illustration, Schubbach compare Deep Blue, l'ordinateur qui a battu Kasparov aux échecs en 1997, à AlphaGo, qui a battu Lee Sedol au go en 2016. Avec Deep Blue, les ingénieurs se targuaient de « pouvoir dire précisément quelles caractéristiques du

280 Sur la différence entre explication et justification dans le cadre du *machine learning,* Arno Schubbach cite O. Biran and K. McKeown, « Human-centric justification of machine learning predictions » in C. Sierra (ed.), *Proceedings of the twenty-sixth international joint conference on artificial intelligence. Main track,* 2017, p. 1461-1467.

système étaient en jeu dans chaque partie[281] ». AlphaGo s'étant calibré sur l'expérience acquise à partir de l'histoire du Go, et sur d'innombrables parties jouées seul, ses développeurs sont incapables d'« expliquer » ses coups.

Sera-t-il possible, à force d'investigations, de trouver une explication satisfaisante au résultat d'un réseau de neurones ? Schubbach passe en revue certaines des méthodes existantes (*feature visualisation, visual explanations…*) et souligne qu'elles ne donnent pas d'explication, mais seulement des justifications. L'explication expose *comment le système arrive à un résultat* tandis que la justification dit *pourquoi le résultat est crédible*. Une explication devrait se référer au calcul effectué et présenter le détail de la procédure. Dans le cas du *deep learning*, nous ne disposons que de justifications qui apportent des éléments pour comprendre pourquoi un résultat peut être considéré comme adéquat, indépendamment de la manière dont il a été calculé.

L'optimisme scientifique nous amène à penser qu'en y apportant suffisamment d'efforts, nous finirons bien par fournir des explications satisfaisantes. Mais Schubbach choisit de considérer l'idée opposée. Quels que soient les moyens déployés, peut-être est-il impossible de reconstituer le cheminement effectué par un algorithme de *deep learning* ? Dans ce cas, le *deep learning* se rapprocherait de la troisième forme de jugement décrite par Kant – le jugement esthétique :

> […] le résultat des opérations de *deep learning* peut être comparé au jugement empirique de Kant si son fonctionnement peut, au moins rétrospectivement, être traduit en règles et critères explicites. Si ce n'est pas possible, alors les résultats semblent plus proches du jugement esthétique[282].

Le jugement esthétique s'effectue sans concept, il trouve ses normes dans la perception individuelle d'œuvres individuelles, sans se traduire en règles générales. En me rendant familier avec un objet spécifique, j'en abstrait un ordre, mais pas de règles qui puisse s'appliquer à d'autres.

En résumé, Arno Schubbach s'appuie sur Kant pour qualifier le type de jugement produit par un algorithme de *deep learning*, et plus précisément le fait que son output n'est pas explicable par une séquence d'arguments s'enchaînant selon des règles logiques. Il rapproche le

281 « I could tell precisely what hardware evaluation features were at play in each game » Feng-hsiung Hsu, *Behind deep blue*, Princeton, Princeton University Press, 2002, p. 200. Cité par Arno Schubbach, *op. cit.* Nous traduisons.

282 « […] the output of the DLN's operations can only be compared to Kant's empirical judgement if its functionality can, at least in retrospect, be tralsated into explicit rules and criteria. It this is not possible, then the outputs appear to be closer to the aesthetic judgment. » Arno Schubbach, *op. cit.* Nous traduisons.

premier type de jugement kantien (*Critique de la raison pure*) à l'intelligence artificielle symbolique, et les jugements empiriques et esthétiques (*Critique de la faculté de juger*) au *deep learning* en se basant sur le fait que les premiers s'expliquent par des règles explicites alors que les seconds se justifient par une familiarité avec l'objet. La « connaissance » y étant encodée par l'expérience, les réseaux de neurones proposent une représentation *empiriste* de la pensée, ce qui leur vaut le surnom de « machines inductives[283] » – « induction » y étant entendue comme équivalent à « connaissance par expérience[284] ». Le mot d'induction ayant plusieurs sens[285], c'est plus précisément celui correspondant à la notion de *généralisation* qui paraît s'appliquer ici : l'« opération par laquelle on étend à toute une classe (généralement indéfinie en extension), ce qui a été observé sur un nombre limité d'individus ou de cas singuliers appartenant à cette classe[286] ». En effet, l'« apprentissage de représentation » effectué par un réseau de neurones consiste à détecter quelles sont les caractéristiques communes aux exemples d'une même classe qui lui ont été soumis, de manière à pouvoir affecter à cette même classe de nouveaux exemples possédant les caractéristiques identifiées.

1.4.3. Les réseaux convolutionnels comme modèle de l'induction

Pour Cameron Buckner[287], la manière dont le *deep learning* fonctionne, et en particulier les réseaux convolutionnels, peut nous éclairer sur le sujet, abondamment débattu tout au long de l'histoire de l'empirisme, de la formation des abstractions à partir d'exemples particuliers. Les variations de contexte, d'orientation, de position, d'éclairage, les superpositions font que nous ne voyons jamais deux fois la même image d'une chose. Dès lors, comment arrivons-nous à reconnaître un objet depuis deux points de vue différents, ou bien que deux objets dissemblables appartiennent à la même catégorie ?

Selon Locke, à force de percevoir des exemples d'une classe, nous identifions un certain nombre de traits communs. Petit à petit, nous devenons capables de distinguer ce qui est

283 C'est ainsi que les qualifient Cardon et al., *op. cit.*
284 André Lalande, *Vocabulaire technique et critique de la philosophie,* Paris, PUF / Quadrige, 2016, p. 506. Le dictionnaire attribue notamment à Leibniz cette manière d'utiliser le mot « induction ».
285 *Ibid,* p. 506-508.
286 *Ibid,* p. 381.
287 Cameron Buckner, « Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks », *Synthese,* 12, p. 1-34, 2018.

nécessaire de ce qui est accidentel. Nous formons une représentation de l'objet dépouillée de caractéristiques non pertinentes – l'essence de la chose. De ce point de vue, une abstraction est l'ensemble des caractéristiques essentielles d'une chose. On reconnaît la description de l'induction par Aristote, que Gauker appelle **l'abstraction-par-soustraction**[288]. Mais Locke n'explique pas comment s'opère la distinction entre nécessaire et accidentel. Surtout, nous semblons avoir une pratique plus souple de l'induction. Nous n'avons aucun mal à considérer comme appartenant à la même classe deux objets dont les qualités nécessaires sont mutuellement incompatibles. Nous arrivons également à reconnaître un objet sans que soient visibles ces caractéristiques essentielles : je peux reconnaître une chaise depuis le ciel, sans voir ses quatre pieds.

Selon Berkeley et Hume, nous sélectionnons ou inventons une série d'exemples appropriés capables de représenter toute la classe et ses différentes sous-catégories potentiellement inconsistantes – méthode que Gauker appelle **l'abstraction-par-représentation.** De ce point de vue, une abstraction est une collection d'exemples suffisamment représentatifs pour permettre de détecter une ressemblance avec les objets de la classe. Mais le fonctionnement de ce type d'abstraction reste mystérieux : comment faisons-nous pour sélectionner ou inventer le *bon* exemple, celui qui est le plus représentatif des éléments de la classe ?

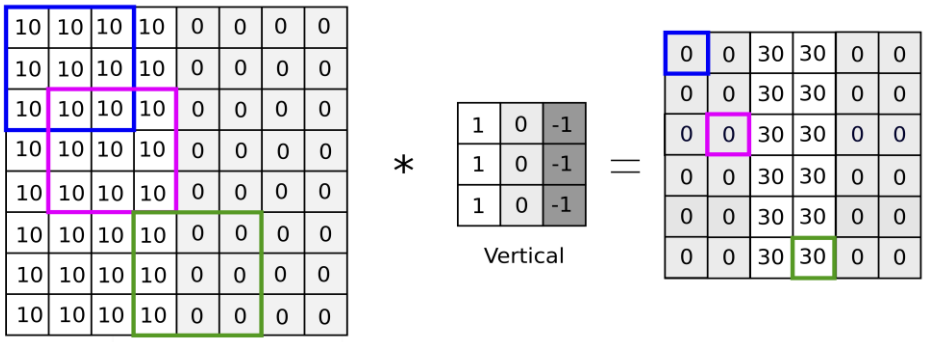
Les deux types d'abstraction sont problématiques. Avec l'abstraction-par-soustraction, nous ignorons comment s'effectue le « trajet mental » des exemples particuliers vers les catégories abstraites : par quel procédé l'esprit arrive-t-il à distinguer les caractéristiques pertinentes de celles à négliger ? Avec l'abstraction-par-représentation, c'est le « trajet mental » depuis les catégories abstraites vers les exemples particuliers qui pose problème : comment l'esprit fait-il pour sélectionner ou produire des représentants pertinents, contenant les bonnes caractéristiques ?

Pour Cameron Buckner, les réseaux convolutionnels profonds combinent les deux mouvements : depuis les exemples particuliers vers la constitution de catégories abstraites, et depuis les catégories abstraites vers la qualification ou l'invention d'exemples particuliers. Ils réalisent à la fois une abstraction-par-soustraction, en dépassant l'écueil de propriétés mutuellement incompatibles, et une abstraction-par-représentation. Il appelle ce type d'abstraction l'**abstraction-par-transformation**.

288 C. Gauker, *Words and images : An essay on the origin of ideas*, Oxford, Oxford University Press, 2011. Cité par Cameron Buckner, *op. cit.*

Afin d'éclairer le propos de Cameron Buckner, revenons en détail sur le fonctionnement des réseaux convolutionnels. On se souvient (voir infra 1.3.10) que c'est l'alternance entre cellules simples et cellules complexes qui leur permet de détecter un objet malgré les variations (changements d'orientation, de taille de luminosité, variations d'échelle, rotations), c'est-à-dire de détecter les caractéristiques nécessaires tout en s'affranchissant de l'excès de détail. Concrètement, les réseaux convolutionnels traduisent cela en trois étapes : les convolutions, la normalisation et le sous-échantillonnage.

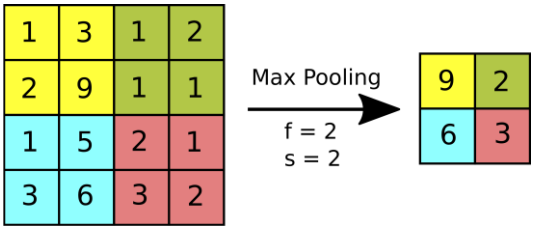
À l'étape de **convolution**, des détecteurs spécialisés (appelés filtres ou *kernels*) parcourent l'ensemble de l'image **à la recherche d'une caractéristique particulière** (par exemple une ligne verticale). Chaque détecteur est une matrice de poids qui, en étant multipliée avec les valeurs d'une zone de l'image, amplifie ce qui est pertinent (ce qui tient d'une ligne horizontale) et minimise ce qui ne l'est pas.



La multiplication des valeurs des différentes zones de l'image (par exemple, les zones bleues, violettes et vertes) par la matrice de poids « vertical » permet de mettre en évidence la présence d'une ligne verticale au centre. Source : Anh H. Reynolds, « Convolutional Neural Networks (CNNs) », https://anhreynolds.com/blogs/cnn.html page consultée le 10 novembre 2020.

L'opération de convolution permet donc **d'amplifier certains aspects de l'image pour faire apparaître la caractéristique recherchée**, en préservant ses caractéristiques (position, orientation, etc.).

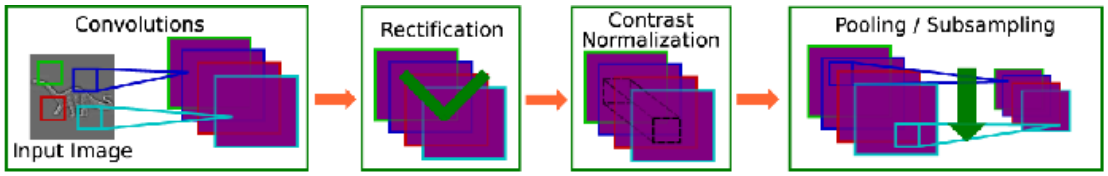
Après une étape de **normalisation du signal**, celui-ci est transmis à l'étape finale, appelée *pooling* ou *subsampling,* dont le rôle est d'en **réduire le volume** (*downsampling* ou sous-échantillonage). Concrètement, le signal issu de l'image est divisé en petites zones dont on prend la moyenne (*average pooling*) ou le maximum (*max pooling*).



Exemple de max-pooling : seule la valeur maximum de chaque zone de l'image est retenue. Source : Anh H. Reynolds, « Convolutional Neural Networks (CNNs) », https://anhreynolds.com/blogs/cnn.html page consultée le 10 novembre 2020.

Cette étape condense l'information de manière à **ne retenir que l'essentiel**, en l'occurence la présence d'une caractéristique donnée (ligne verticale) indépendamment de sa localisation précise et de son orientation. Si elle est « à peu près » présente, cela suffit.

La succession des trois étapes permet de reproduire l'alternance entre cellules simples qui détectent la caractéristique et passent le signal en préservant ses spécificités (ici jouées par l'étape de convolution) et cellules complexes (ici, la normalisation et le sous-échantillonage) qui permettent de préserver l'essentiel du signal (« il y a une ligne verticale ») sans prendre en compte trop de détails. La première opération est linéaire, elle **conserve les specificités de la caractéristique détectée**, tandis que les suivantes sont non-linéaires, elles permettent une **tolérance aux variations**.



À l'étape de convolution, quatre « filtres » scannent l'image à la recherche de caractéristiques particulières. Les quatre résultats sont normalisés (étapes de « rectification » et « contrast normalization ») puis sous-échantillonés (« pooling / subsampling »). Image : Lecun, Yann, Koray Kavukcuoglu, Clement Farabet, « Convolutional Networks and Applications in Vision », op. cit.

Une fois que ces trois étapes ont permis de détecter et d'« abstraire » des caractéristiques élémentaires, le processus est répété de manière à détecter des caractéristiques plus élaborées (combinant les caractéristiques élémentaires). Puis le processus est réitéré jusqu'à une dernière couche où les caractéristiques identifiées permettent de classer l'objet. Par exemple, des oreilles pointues et une moustache permettent de reconnaître un chat. À chaque stade, les « filtres » détectent les caractéristiques agrégées, puis le signal est normalisé et sous-échantillonné. Ainsi,

l'alternance entre convolution et normalisation / sous-échantillonnage permet à chaque fois de détecter les caractéristiques tout en s'affranchissant des variations non pertinentes.

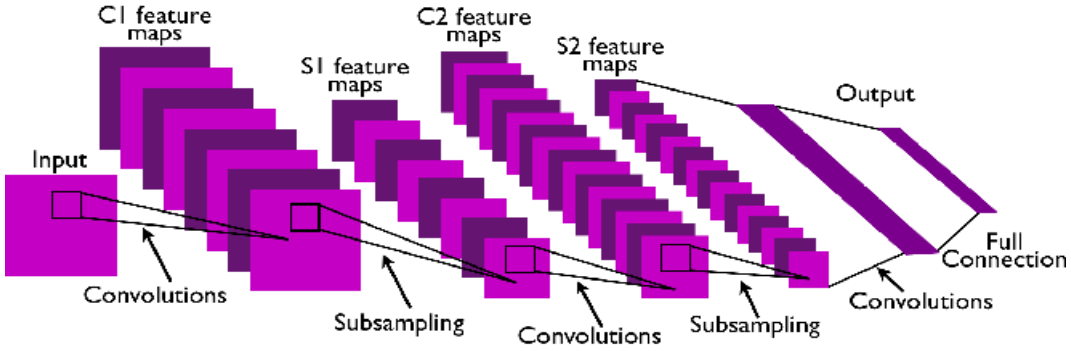


Illustration de l'alternance entre convolutions et normalisation / sous-échantillonnage amenant de l'image (input) à la classification (output). Image : Lecun, Yann, Koray Kavukcuoglu, Clement Farabet, « Convolutional Networks and Applications in Vision », op. cit.

C'est donc le fait de réitérer l'alternance entre les deux opérations qui permet de préserver et d'accentuer les caractéristiques pertinentes tout en s'affranchissant de leurs variations non pertinentes. À travers le processus d'apprentissage, le réseau acquiert les poids qui le rendent sensible à certains ensembles de caractéristiques tout en le rendant indifférent à certaines variations. Pour Cameron Buckner, un réseau de neurones convolutionnel

> implémente une forme d'abstraction hiérarchique qui réduit la complexité de l'espace des caractéristiques d'un problème (et évite le surapprentissage [*overfitting*] sur les exemples d'entraînement) en le transformant itérativement en un format de représentation qui préserve et accentue les caractéristiques pertinentes tout en écartant les variations parasites[289].

Ainsi, les réseaux convolutionnels constituent une solution pertinente à l'un des problèmes classiques posés par la notion d'induction, à savoir comment sont détectés les traits communs définissant une classe avec assez de souplesse pour s'adapter à la variabilité des individus et des situations. Bien que nous ne sachions pas si les animaux et les humains procèdent à la

289 « implement a form of hierarchical abstraction that reduces the complexity of a problem feature space (and avoids overfitting the network's training samples) by iteratively transforming it into a simplified representational format that preserves and accentuates task-relevant features while controlling for nuisance variations. » Cameron Buckner, *op. cit.* Nous traduisons.

manière des réseaux convolutionnels, la solution proposée est pertinente. Elle a le mérite de fonctionner : les réseaux convolutionnels sont capables d'assigner des objets hétérogènes, malgré une grande variabilité de situations, aux bonnes classes.

Au fil du réseau, en passant les différentes couches, le signal initial (la matrice de valeurs des pixels de l'image) est transformé d'étapes en étapes en une représentation de l'objet de plus en plus tolérante aux variations et de plus en plus apte à être assignée à une classe – c'est-à-dire de plus en plus proche de l'abstraction à laquelle il correspond. Ainsi, se pencher sur les dernières étapes du processus devrait permettre d'esquisser une notion concrète de ce qu'est une abstraction.

Au fil du réseau, le signal originel (les pixels de l'image) a été transformé de manière à aboutir à **une série de valeurs quantifiant la présence de chaque caractéristique discriminante**. Au niveau des dernières couches du réseau, c'est la combinaison des caractéristiques complexes (détectées par agrégation des caractéristiques élémentaires) qui déclenche la classification. L'objet perçu y est donc **encodé sous la forme d'un vecteur** contenant autant de dimensions qu'il y a de caractéristiques discriminantes. Si le réseau fait bien son travail d'apprentissage, l'encodage réalisé permet de **regrouper les objets similaires dans l'espace vectoriel des caractéristiques discriminantes**. Dès lors, on peut considérer que, pour un réseau de neurones convolutionnel, **une abstraction est une région de cet espace vectoriel des caractéristiques** – ouvrant la voie à une notion mathématique de ce qu'est une abstraction, voire de l'imagination.

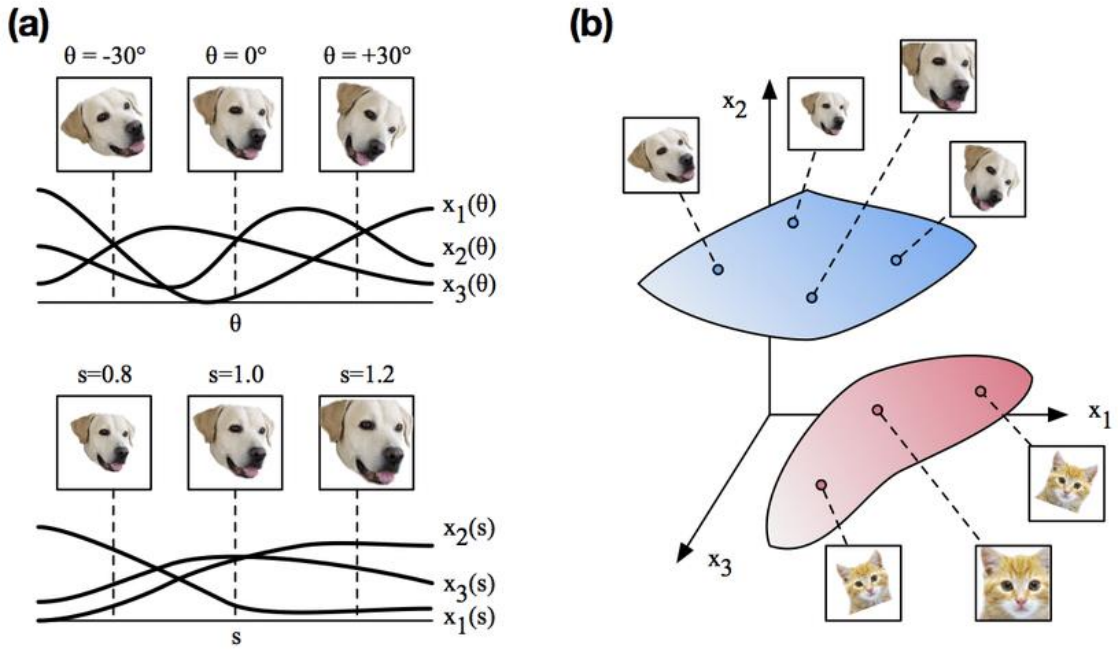
1.4.4. L'abstraction comme région de l'espace vectoriel des caractéristiques

Du point de vue du réseau de neurone, chaque exemple n'est rien d'autre qu'un vecteur encodant sa valeur pour chacune des « dimensions », c'est-à-dire des caractéristiques prises en compte[290]. Tout se passe comme si les exemples étaient positionnés dans un espace vectoriel qui contient autant de dimensions qu'il y a de caractéristiques discriminantes. On appelle cet espace vectoriel *l'espace latent (latent space)*. La classification qu'effectue le réseau de

290 « Perceptual similarity is a multi-dimensional vector space – with each dimension standing for a perceptually discriminable feature – that plots an agent's perceptuel experience of each exemplar to a unique vector. Vector distance in this space marks the degree of perceived similarity between the different exemplars. » Cameron Buckner, *op. cit.* p. 11.

neurones revient à y tracer des frontières délimitant des collections d'objets similaires – des catégories. On appelle ces collections des « manifolds »[291].

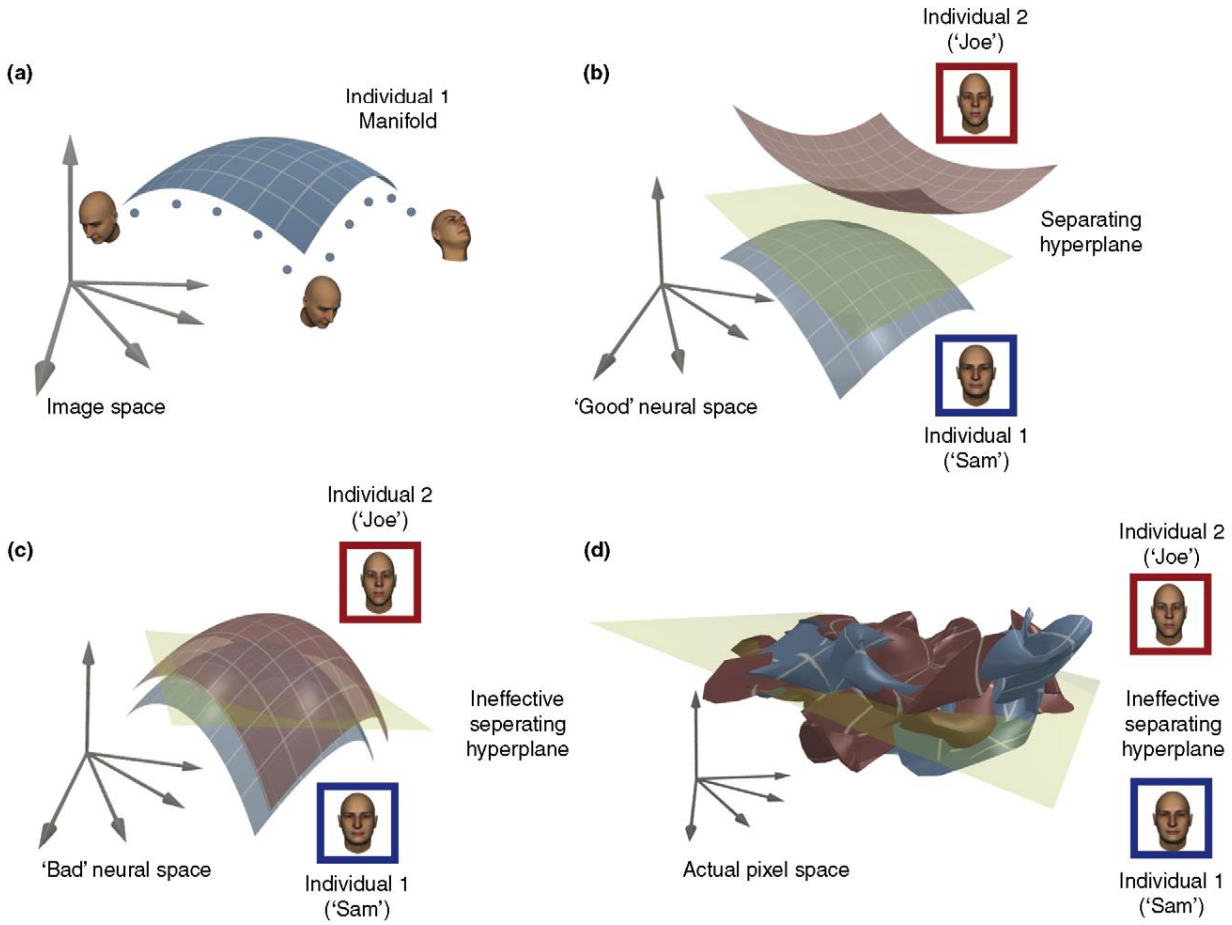
Pour résumer, on peut considérer que le réseau assigne une « adresse » à chaque objet et regroupe dans une même catégorie les objets dont les « adresses » sont proches.



Les graphiques de gauche représentent la fréquence d'activation de trois neurones en réponse à une image de chien dont on fait varier l'orientation θ (en haut) et l'échelle s (en bas). La réponse pour chaque image particulière (orientation et échelle) correspond à un vecteur à trois dimensions. À droite, la réponse des neurones aux variations d'images forme une collection dans l'espace à trois dimensions. La frontière tracée autour de cette collection définit la catégorie « chien » (zone bleue). D'autres objets dont on aura fait varier l'image forment autant d'autres collections dans l'espace vectoriel (ici, la zone rouge correspondant aux images de chats). Image : Sueyeon Chung, Daniel Lee, Haim Sompolinsky, « Classification and Geometry of General Perceptual Manifolds » Physical Review, *X 8 (3), 2017.*

La difficulté de l'opération de classification réside dans le fait que, dans l'espace vectoriel, les différentes collections sont enchevêtrées. Les multiples variations d'un objet amènent la collection à se superposer avec les autres collections.

291 « A 'manifold' in perceptual similarity space is a region of this vector space, and category representations can be construed as manifolds. » *Ibid.*



Le graphique (a) représente la collection d'images correspondant aux portraits de l'individu 1. Dans le cas du graphique (b), les portraits de l'individus 1 et 2 sont séparables par un hyperplan, tandis qu'ils ne le sont pas en (c). Le graphique (d) montre le point de départ de l'opération de classification : les deux collections de portraits sont enchevêtrées. L'enjeu des opérations successives conduites par le réseau de neurones est de trouver des transformations qui, tout en conservant assez d'informations, permettent de passer de l'espace enchevêtré (d) à deux collections séparables (b).

Pour DiCarlo, Zocolan, et Rust, tout se passe comme si les collections ou « manifolds » étaient enchevêtrées, « comme plusieurs feuilles de papier roulées en une seule boule[292] ». Distinguer deux objets – et il y aurait là un point commun entre les opérations effectuées par le cerveau et les réseaux de neurones convolutionnels – revient à trouver une série de transformations de l'espace qui « déplient » la boule de manière à séparer les différentes feuilles[293].

292 DiCarlo, Zoccolan, Rust, « How does the brain solve visual object recognition? », *Neuron*, 9 février, 73 (3), 2012 , p. 415-34.

293 Voir également DiCarlo, James, Cox, « Untangling invariant object recognition. » *Trends in Cognitive Sciences*, 11 2007, p. 333-341.

Ainsi, nous pouvons **définir le processus d'abstraction** comme l'identification des caractéristiques discriminantes d'un objet d'une manière robuste aux variations, couplé à la séparation entre les « feuilles ».

Cela nous permet également de définir ce qu'est une abstraction (selon le modèle des réseaux convolutionnels) : c'est **une région dans l'espace multidimensionnel des caractéristiques** obtenue en transformant celui-ci de manière à ce qu'il y ait la meilleure séparation possible entre les différentes abstractions.

Par contre, il nous est impossible de nous représenter clairement une telle abstraction. Nous pouvons l'illustrer de manière simplifiée en trois dimensions (voir les deux figures précédentes), mais dans les faits, les caractéristiques prises en compte sont trop nombreuses pour que nous puissions les imaginer. La région de l'espace vectoriel correspondant à l'objet « chat » n'est pas à la portée de notre imagination dans la mesure où elle se déploie dans trop de dimensions. L'exemple des réseaux convolutionnels permet ainsi de concevoir comment un cerveau pourrait **encoder une abstraction, sans pour autant que celle-ci soit représentable**.

1.4.5. Mathématiques de l'imagination

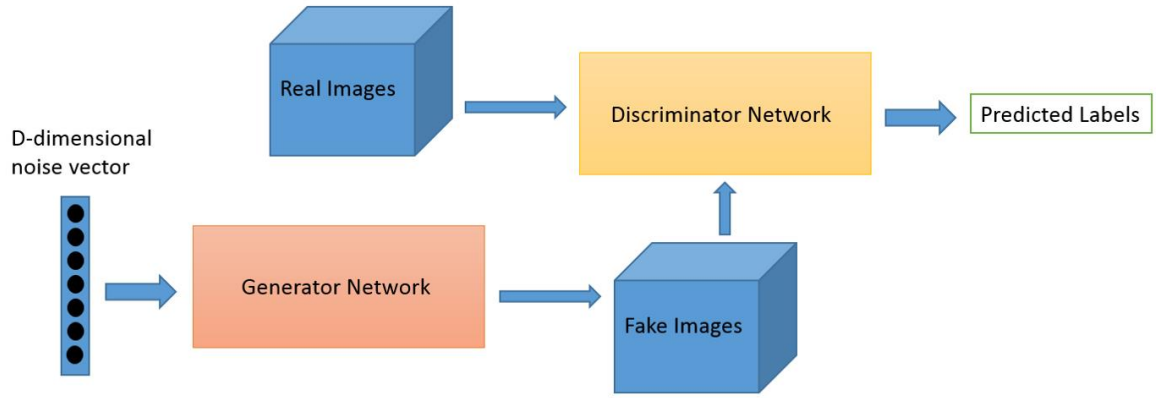
L'abstraction comme région de l'espace vectoriel des caractéristiques permet non seulement de qualifier d'autres objets appartenant à la même catégorie, mais aussi d'en imaginer de nouveaux. Après que le réseau ait vu suffisamment d'images de chats et délimité pour l'objet « chat » une région de l'espace des caractéristiques, je peux y choisir un vecteur (c'est-à-dire, donner une valeur à chacune des caractéristiques), puis le ré-échantillonner de manière à obtenir une image crédible de chat ne correspondant à aucun chat réel.

Le ré-échantillonnage peut se faire selon différentes méthodes. En 2014, Ian Goodfellow et ses collègues proposent la méthode des réseaux adversariels génératifs ou GAN (*generative adversarial networks*)[294]. Un premier réseau de neurones convolutionnels classique (appelé le discriminateur) est entraîné à classer des images – par exemple, à distinguer les images de chat des autres. Un deuxième réseau de neurones (appelé le générateur), paramétré à partir de bruit, génère des images aléatoires qu'il soumet au jugement du premier. Le deuxième

294 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, « Generative Adversarial Networks », arXiv:1406.2661, 2014.

réseau utilise la classification du premier comme *feedback* pour réajuster ses paramètres de manière à produire des images qui puissent passer pour des images de chat.

Les GANs fonctionnent comme une collaboration entre un expert et un faussaire. À force de multiplier les essais, le faussaire finit par trouver les critères que l'image doit satisfaire pour être qualifiée par l'expert. Il devient capable de fabriquer des faux que l'expert ne distingue plus des vrais.

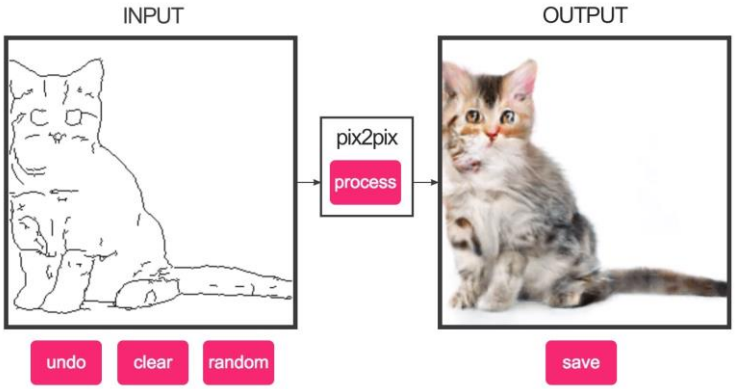


Le jugement du réseau de neurones « discriminateur » (entraîné à partir d'images « réelles ») est utilisé par le réseau « générateur » (comme taux d'erreur à minimiser) pour calibrer ses productions. Source : Jon Bruner and Adit Deshpande, « Generative Adversarial Networks for beginners », O'Reilly, 7 juin 2017, https://www.oreilly.com/content/generative-adversarial-networks-for-beginners/

L'*input* du générateur est un vecteur pris au hasard dans l'espace latent. Le *feedback* du discriminateur contraint le générateur à situer les vecteurs au bon endroit, dans la région de l'espace regroupant les exemples de chats, et à les rééchantilloner de manière à ce que les détails ajoutés fassent une image de chat crédible (qui ne sont pas des reproductions des chats de la base de données d'exemples[295]).

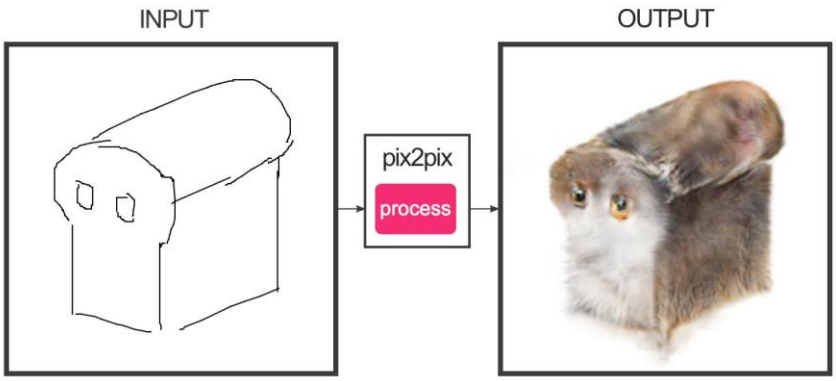
Les GANs constituent un mécanisme à même de fabriquer des images pertinentes d'une catégorie donnée. La production par le générateur de nouvelles images de chat implique qu'il ait correctement intégré une notion pertinente de chat (ou tout au moins de « image de chat »).

295 Goodfellow et al. Insistent bien sur ce point, *op. cit.*



L'application pix2pix permet, à partir un dessin, de produire une image de chat. Source : Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, « Image-to-Image Translation with Conditional Adversarial Networks », arXiv:1611.07004, 2017.

Nous disposons ici d'une forme **d'imagination artificielle**[296], avec la même liberté que notre imagination. Il est possible de donner au chat une forme incongrue, que le réseau va ensuite ré-échantillonner *comme s'il s'agissait d'un chat.* Il est ainsi possible de lui faire produire une image chat en forme de brioche, ou toute autre forme.



Chat en forme de brioche. Source : Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, « Image-to-Image Translation with Conditional Adversarial Networks », op. cit.

296 Le premier à utiliser le terme d'« imagination artificielle » est Arnold Kaufmann, il le présente comme les différentes manière, pour des humains, d'utiliser les ordinateurs pour explorer les « morphologies ». « L'imagination artificielle (heuristique automatique) », *Revue française d'informatique et de recherche opérationnelle*, tome 3, n°3, 1969, p. 5-24

Autres exemples de chats de forme variée. Source : Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, « Image-to-Image Translation with Conditional Adversarial Networks », op. cit.

Les programmes les plus récents (Dall-E, Midjourney, Stable Diffusion…) sont capables, à partir d'une simple description textuelle (appelé le *prompt*), d'élaborer entièrement une image. À titre d'exemple, voici ce que produisent Stable Diffusion et Dall-E 2 à partir de la description « A Boston Terrier jedi holding a dark green lightsaber, photorealistic » :

Image : Nir Barazida, « Stable Diffusion: Best Open Source Version of DALL·E 2 », Towards Data Science, *30 août 2022,* https://towardsdatascience.com/stable-diffusion-best-open-source-version-of-dall-e-2-ebcdf1cb64bc page consultée le 3 septembre 2022

Les programmes de production d'image (et de vidéo[297]) produisent à l'envi et nous font entrer dans une ère de prolifération des formes. De nombreux artistes se saisissent des GANs[298], puis des outils plus récents, déclenchant une polémique sur un éventuelle « fin du métier d'artiste[299] ». Nous nous dirigeons vers la situation anticipée par Grégory Chatonsky : sur le site *Capture,* il imagine un groupe de rock qui « produit chaque heure de nouvelles musiques, paroles, images, vidéos et produits dérivés ». Le groupe est « si productif que personne ne peut tout consommer[300] ».

Dans la mesure où la région de l'espace vectoriel délimitant une classe est effectuée malgré les variations s'appliquant à un objet, elle encode aussi ses variations. Par exemple, une

297 En septembre 2022, Meta annonce *Make-A-Video* et Google annonce *Imagen Video,* des programmes produisant des vidéos courtes à partir de *prompts.*

298 Par exemple le collectif Obvious Art, qui baptise une série d'oeuvres « Portraits de Belamy » en hommage (et jeu de mot) au créateur des GANs, Ian Goodfellow. https://obvious-art.com/portfolio/edmond-de-belamy/ page consultée le 20 novembre 2020.

299 Voir section 3.4.8. Personne à qui parler

300 Grégory Chatonsky, « Capture », http://chatonsky.net/capture/ page consultée le 20 novembre 2020.

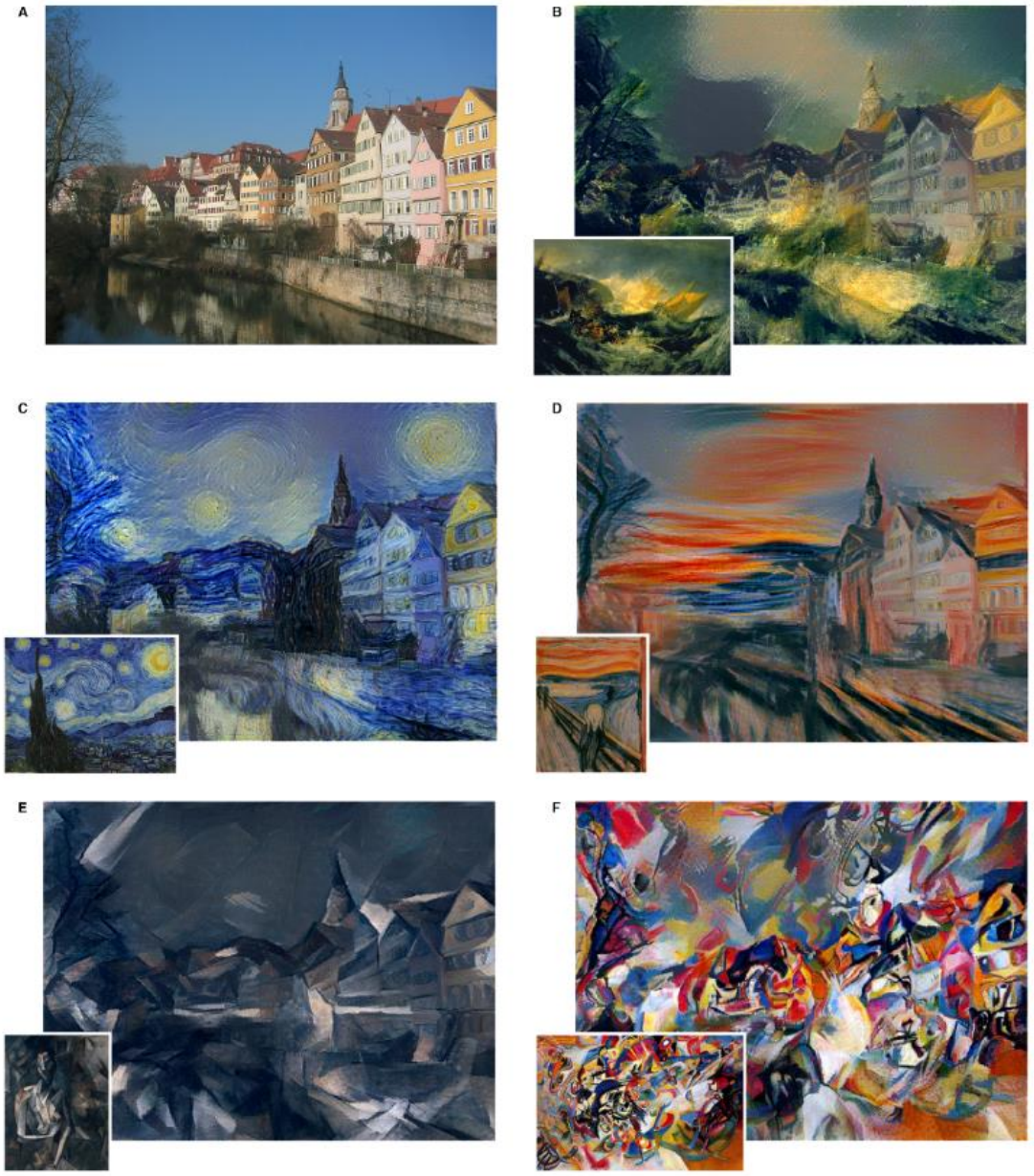
fois que le réseau a été entraîné à reconnaître qu'il s'agit du même objet lorsqu'il est tourné vers la droite ou vers la gauche, il suffit de soustraire le vecteur « objet tourné vers la droite » à celui de l'« objet tourné vers la gauche » pour avoir un vecteur correspondant à la variation « rotation de la droite à la gauche ». L'espace latent encode ce qui est commun à deux individus, indépendamment de leur participation à la même classe. Autrement dit, l'espace latent **encode les qualités, et plus largement les analogies**.

En faisant de l'arithmétique avec les vecteurs correspondant aux objets encodés dans l'espace latent, il est possible « trouver des directions intéressantes », c'est-à-dire d'abstraire des qualités et des analogies indépendamment des objets, sous la forme de vecteurs.

En soustrayant le vecteur de l'image « homme sans lunettes » à celui de « homme avec lunette », on isole le vecteur « avoir des lunettes » que l'on peut ajouter à une image de femme sans lunettes, de manière à obtenir au final une image de femme avec des lunettes. L'opération met en évidence que l'espace latent encode bien le vecteur « avoir des lunettes ». Source : Alec Radford, Luke Metz, Soumith Chintala, « Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks », arXiv:1511.06434, 2016.

L'image source (à gauche) est modifiée par translation dans l'espace vectoriel le long des vecteurs correspondant à la pose, l'âge, l'expression, le port de lunette. Source : Yujun Shen, Jinjin Gu, Xiaoou Tang, Bolei Zhou, « Interpreting the Latent Space of GANs for Semantic Face Editing », arXiv:1907.10786, 2020.

L’espace latent n’encode donc pas que les objets, mais aussi leurs variations. À partir du moment où une variation peut s’appliquer à plusieurs objets, il est possible d’abstraire cette variation sous la forme d’un vecteur et de l’appliquer à un nouvel objet. Et cela vaut pour le style : si l’on dispose de plusieurs œuvres du même auteur, il est possible d’abstraire le style des images pour l’appliquer à de nouvelles images. Les réseaux de neurones sont une formidable machine à fabriquer des pastiches.



L’image d’entrée (A) est transformée en lui appliquant le « style » de différents peintres : Turner (B), Van Gogh (C), Munch (D), Picasso (E), Kandinsky (F). Source : Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, « A Neural Algorithm of Artistic Style », arXiv:1508.06576, 2015.

En combinant la détection du style opérée par les réseaux de neurones avec la précision de l'impression 3D, une équipe de ING et de Microsoft entreprend de créer un tableau de Rembrandt perpétuant son style, en allant dans le détail de chaque coup de pinceau.

Image : The Next Rembrandt, ING Group, Wikimedia Commons, https://fr.wikipedia.org/wiki/Fichier:The_Next_Rembrandt_1.jpg page consultée le 20 novembre 2020.

Les réseaux convolutionnels sont donc capables d'attribuer la bonne classe aux objets et de dissocier les qualités de l'objet pour les transférer à un autre objet. Pour Rodney Brooks, ainsi que le rappelle Cameron Buckner[301], cette faculté d'abstraction est l'essence de l'intelligence. En 1991, il annonçait que les programmes seraient réellement intelligents lorsqu'ils deviendraient capables, par exemple, de reconnaître une chaise arbitraire dans une image tout aussi arbitraire[302]. Pour pouvoir affirmer que les réseaux de neurones fournissent la clef de l'intelligence, il faudrait que leur faculté d'abstraction s'étende au-delà des images. Leur succès au jeu de Go, bien qu'il passe par le traitement des images (l'analyse du *goban*), implique l'abstraction de notions de stratégie comme l'influence, les connections, les stabilités – indépendamment de parties spécifiques.

301 Cameron Buckner, « Empiricism without Magic : Transformational Abstraction in Deep Convolutional Networks », *op. cit.*

302 « In particular the concept of what is a chair is hard to characterize simply. There is certainly no AI vision program which can find arbitrary chairs in arbitrary images... » Rodney Brooks, « Intelligence without representation », *Artificial Intelligence,* 47(1-3), 1991, p. 139-159.

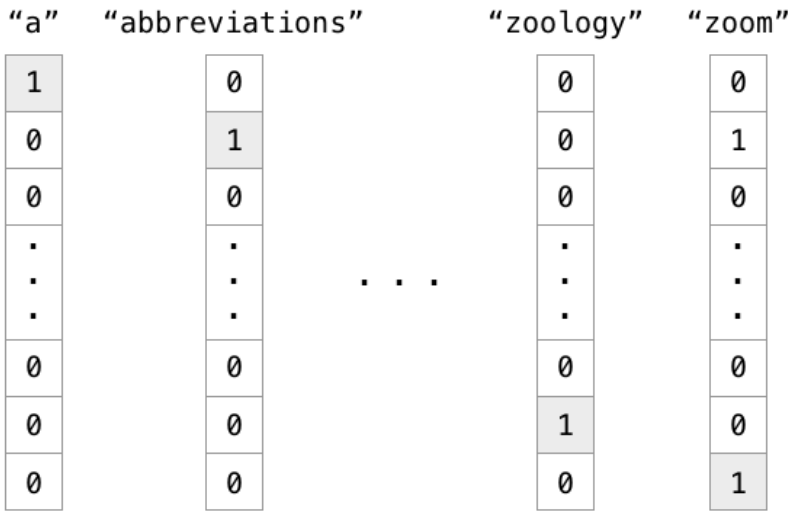
Pour évaluer la capacité des réseaux de neurones à effectuer des abstractions dans un domaine plus théorique que le traitement des images, nous allons nous pencher maintenant sur leur manière de traiter le langage.

1.4.6. Une représentation mathématique du langage

Le *deep learning* est arrivé plus tardivement dans le domaine du traitement automatique du langage (TAL, ou NLP pour *natural langage processing*) car les modèles précédents avaient déjà de bonnes performances. Un tournant s'opère en 2013, avec les articles de Mikolov et de ses collègues présentant *word2vec*, une nouvelle manière de représenter les mots sous la forme de vecteurs[303]. *Word2vec*, puis d'autres outils de la même famille, deviennent les constituants de base des algorithmes de traitement automatique du langage et déclenchent une mutation du champ aussi spectaculaire que pour le domaine du traitement de l'image.

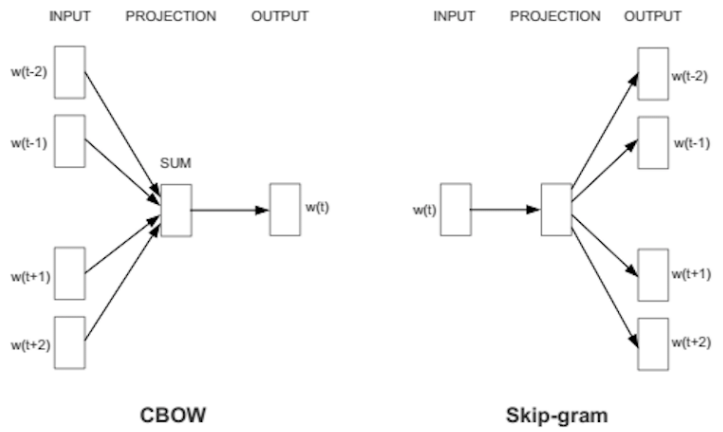
Le point de départ de *word2vec* est un encodage de chaque mot sous la forme de vecteurs contenant autant de dimensions qu'il y a de mots dans le dictionnaire, d'une valeur de 1 dans la dimension correspondant à sa place dans le dictionnaire, et de zéro pour toutes les autres dimensions (appelés *one hot vectors*). Par exemple, si le mot « abbreviations » est le deuxième mot d'un dictionnaire anglais de dix mille mots, il est représenté par un vecteur de dimension dix mille, de valeur 1 en deuxième position et zéro ailleurs.

303 Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, « Efficient Estimation of Word Representations in Vector Space », arXiv:1301.3781, 2013 et Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, « Distributed Representations of Words and Phrases and their Compositionality », arXiv:1310.4546, 2013. La vectorisation de mots en tant que telle n'est pas nouvelle, elle est utilisée dans des moteurs de recherches depuis des dizaines d'années.



Liste des mots du dictionnaire sous la forme de vecteurs de valeur 1 dans la dimension correspondant à leur place et zéro dans les autres dimensions. Source : Huang (Steeve) Kung-Hsiang, « Word2Vec and FastText Word Embedding with Gensim », Towards data science, *4 février 2018, https://towardsdatascience.com/word-embedding-with-word2vec-and-fasttext-a209c1d3e12c , page consultée le 20 septembre 2020.*

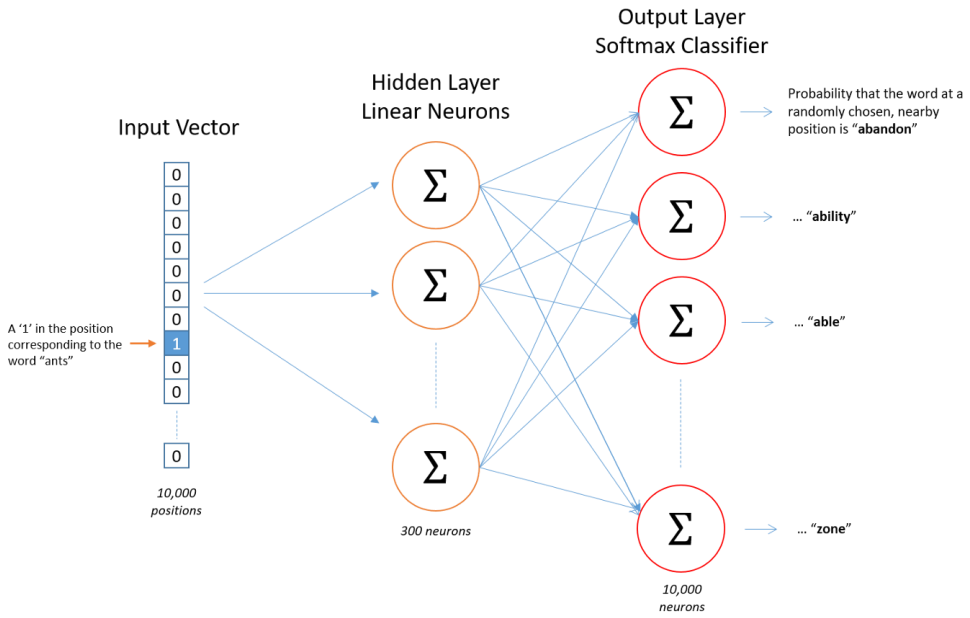
L’article de Mikolov et de ses collègues propose deux manières d’entraîner un réseau de neurones. Avec la méthode CBOW, le réseau de neurones apprend à prédire un mot en fonction de son contexte (comme un texte à trous), tandis qu’avec la méthode *skip-gram*, le réseau de neurones apprend à proposer un contexte à partir d’un mot.



Avec la méthode CBOW, le réseau de neurones reçoit quatre vecteurs de mots en entrée (le « contexte », par exemple « Je », « me », « un » et « café ») et doit prédire un vecteur de mot en sortie (par exemple « verse »). À l’inverse, avec la méthode Skip-gram le réseau de neurones reçoit un vecteur de mot en entrée et doit prédire les quatre mots qui l’entourent (le « contexte »). Image : Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, « Efficient Estimation of Word Representations in Vector Space », arXiv:1301.3781, 2013.

Comme pour les exemples exposés précédemment, le réseau de neurones se calibre par essai-erreur sur un grand nombre d'exemples prédéfinis – le corpus d'apprentissage. Par exemple, il reçoit quatre mots de « contexte » et il propose un mot pour remplir le « trou ». La correction lui fournit un taux d'erreur qu'il cherche à minimiser en modifiant ses paramètres de manière à transformer ces vecteurs pour qu'ils intègrent les relations entre les mots telles qu'elles apparaissent dans un corpus d'exemples. S'il répond « chien » au lieu de « chat », l'erreur est moindre que s'il avait répondu « étagère », aussi retient-il qu'il y a proximité de contexte entre « chien » et « chat ».

Les méthodes CBOW et *Skip-gram* ne sont pas du *deep learning*. Il suffit de réseaux de neurones très simples ne contenant qu'une seule couche intermédiaire. Ils reçoivent en entrée le ou les *one hot vectors* et produisent en sortie le ou les mots compagnons les plus probables.



Exemple de réseau de neurone de type Skip-gram. *Le vecteur* one hot *correspondant au mot* ants *est traité par une couche intermédiaire de trois cents neurones de manière à lui associer, pour chacun des dix mille mots autres mots du vocabulaire, la probabilité qu'il apparaisse dans son « contexte ». Image : Manish Chablani, « Word2Vec (skip-gram model): PART 1 - Intuition »,* Towards data science, *14 juin 2017, https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b page consultée le 20 novembre 2020.*

La sortie d'un réseau de neurones *Skip-gram* donne la probabilité, pour chaque mot du vocabulaire, qu'il apparaisse dans le « contexte » du mot reçu en entrée. Si, grâce à un corpus d'apprentissage suffisant, le réseau de neurones est bien calibré et fonctionne, cela implique que la couche intermédiaire de trois cents neurones soit arrivée à une combinaison de poids et

de biais qui contient l'information nécessaire à la tâche d'attribuer un contexte crédible à chaque mot : elle encode des informations significatives au sujet des relations entre les mots du vocabulaire.

Au niveau de la couche intermédiaire, chaque mot est encodé (en anglais, *embedded*) sous la forme d'un vecteur de dimension trois cent (d'où le terme de *word embedding*)[304]. Puisque cela permet de mener à bien la tâche d'association entre un mot et son contexte, cet ensemble de vecteurs condense l'information au sujet des relations de proximité entre mots du vocabulaire. C'est une mise en pratique de l'hypothèse distributionnelle, énoncée par Firth en 1957, selon laquelle le sens d'un mot est donné par son entourage[305]. En associant les mots et leurs contextes usuels, le réseau apprend à représenter les mots en vecteurs d'une manière telle que deux mots « similaires » donnent deux vecteurs « proches ». Tout repose sur l'hypothèse que deux mots sont similaires s'ils peuvent être utilisés dans le même contexte. Autrement dit, les distances entre mots peuvent être approchées par les différences de contexte d'usage. Les vecteurs correspondant à « chat » et « chien » seront plus proches, dans l'espace vectoriel correspondant (ils apparaissent dans des contextes similaires), que le mot « étagère », et il est possible de mesurer leur distance. Par exemple, la liste des dix mots les plus proches de « house » est « houses », « bungalow », « apartment », « bedroom », « townhouse », « residence », « mansion », « farmhouse », « duplex », et « homes »[306]. Il ne s'agit pas de la notion de synonyme, mais plutôt d'équivalence, ou de subsituabilité.

Disposer des mots sous la forme de vecteurs, c'est en avoir une représentation mathématique et donc rendre possible un certain nombre d'opérations : mesurer la distance, soustraire, ajouter, multiplier… Etant donné la manière dont le réseau est entraîné, la distance mesurée est significative, elle représente le degré de similarité (ou plus précisément, de substituabilité) entre les mots. Une projection de l'espace vectoriel sur deux ou trois dimensions permet de rendre visible à l'œil humain que la répartition des mots se fait selon des relations de proximités (de sens, d'usage, de fonction…)[307].

L'addition entre deux mots-vecteurs revient à approcher les mots qui apparaissent au croisement de leurs deux contextes d'usage, ce qui donne des résultats intéressants.

304 De la même manière que la dernière couche d'un réseau convolutionnel « encode » les images dans l'espace vectoriel des caractéristiques discriminantes (l' « espace latent »). Voir supra.

305 « You shall know a word by the company it keeps » J.R. Firth, « A synopsis of linguistic theory 1930–1955 », *Studies in linguistic analysis,* Oxford, Blackwell, 1957, p. 1–32.

306 Nous reprenons l'exemple donné par Juan-Luis Gastaldi « Why Can Computers Understand Natural Language ?: The Structuralist Image of Language Behind Word Embeddings », *Philosophy & Technology,* 34, 2021.

307 Pour un outil d'exploration de projection des mots en trois dimensions, voir http://projector.tensorflow.org/

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

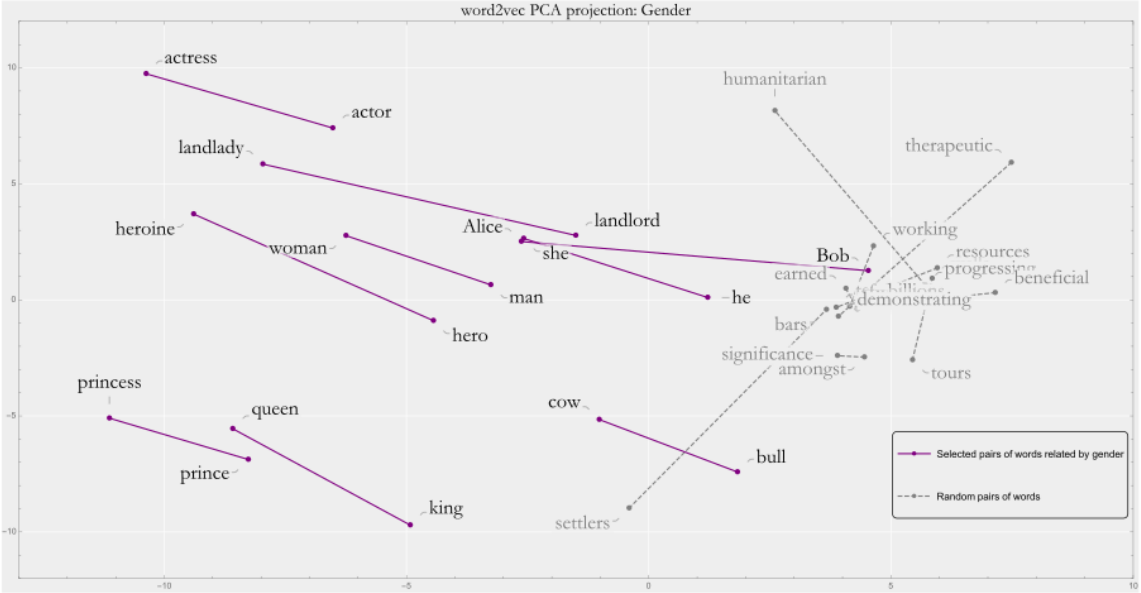
Si on additionne les vecteurs « French » et « actress » les quatre vecteurs les plus proches du résultat sont « Juliette Binoche », « Vanessa Paradis », « Charlotte Gainsbourg » et « Cecile De ». Illustration, Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, « Distributed Representations of Words and Phrases and their Compositionality », arXiv:1310.4546, 2013.

Plus surprenant, Mikolov et son équipe ont constaté que la combinaison d'additions et de soustractions permet de mettre en évidence certaines relations structurantes. Ainsi, en soustrayant le vecteur « Man » au vecteur « King », et en y ajoutant le vecteur « Woman », ils obtiennent un vecteur proche de « Queen ». Autrement dit

$$King - Man + Woman = Queen$$

Le fait de soustraire « Man » et d'ajouter « Woman » permet de passer d'un genre à un autre : de King à Queen mais aussi de « prince » à « princess », « landlord » à « landlady », « bull » à « cow », et ainsi de suite – avec des directions similaires comme si le même vecteur correspondant à la relation « genre » sous-tendait chacune de ces transformations. Autrement dit, l'encodage des relations entre les mots dans l'espace vectoriel (le *word embedding*) permet d'effectuer des **raisonnements analogiques :** ce que donne pour « bull » l'application de la relation qui va de « prince » à « princesse » est « cow ».

Le dispositif de *word embedding* va donc plus loin que la mise en pratique de l'hypothèse distributionnelle puisqu'il ne capture pas seulement la similarité (ou la substituabilité) des mots, mais aussi les relations d'analogie.



Une projection de l'espace vectoriel des mots sur un plan en deux dimensions (via une analyse par composantes principales) permet de visualiser le vecteur encodant la relation de genre. Illustration : Juan Luis Gastaldi, op.cit.

Le vecteur de genre étant inféré par les co-occurrences des mots du corpus d'entraînement, cela permet de mettre en évidence le sexisme dudit corpus. Appliquée à certains mots, la translation du masculin vers le féminin amène à des mots plus péjoratifs. Par exemple, elle donne « nurse » comme le « féminin » de « doctor » si le corpus d'entraînement ne présente pas ou peu d'occurences de médecins féminins, ni d'infirmiers masculins.

Le genre n'est qu'une des nombreuses relations entre les mots que Mikolov et son équipe ont été surpris de déceler, par le biais de diverses opérations élémentaires, dans l'espace vectoriel des mots. Ils ont également mis en évidence le vecteur allant d'un pays à sa capitale, d'une personne célèbre à sa profession, d'un élément chimique à son abréviation, d'un pays à son plat le plus typique, et ainsi de suite. À ces éléments géographiques, historiques et culturels, s'ajoutent des éléments proprement linguistiques : le vecteur allant du singulier au pluriel, d'un mot à son superlatif, à son antonyme, d'un verbe au présent à sa forme au passé...etc.

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Illustration, Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, « Efficient Estimation of Word Representations in Vector Space », arXiv:1301.3781, 2013.

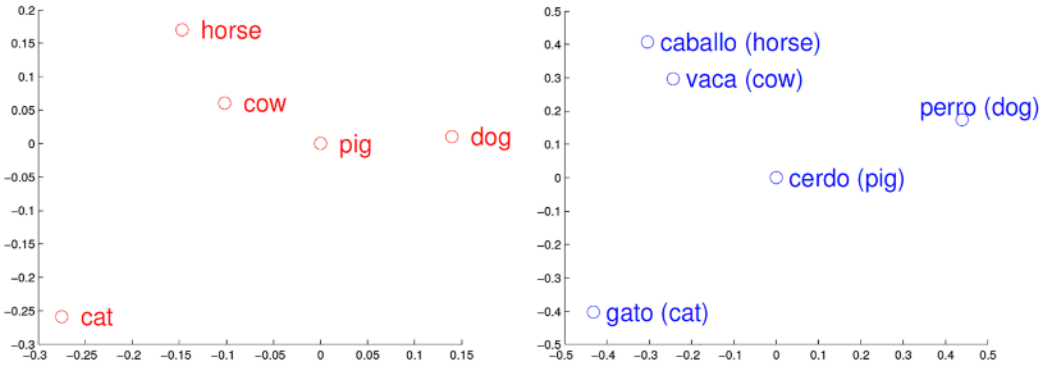
Des travaux ultérieurs ont permis de montrer que la conjugaison des verbes, réguliers et irréguliers, ainsi que toutes les catégories comparatives, sont également encodées dans l'espace vectoriel[308].

Ainsi, l'encodage du vocabulaire sous la forme de vecteurs capture de nombreuses relations entre les mots, et plus généralement des éléments d'organisation du langage – un fois que les valeurs des vecteurs ont été ajustées par le calibrage du réseau de neurones de manière à encoder leur relation de similarité (ou substituabilité).

Juan-Luis Gastaldi recense les différentes découvertes subséquentes. Plusieurs travaux[309] ont remarqué que le passage d'une langue à l'autre modifie peu la distribution des mots dans l'espace vectoriel, manifestant une certaine invariance.

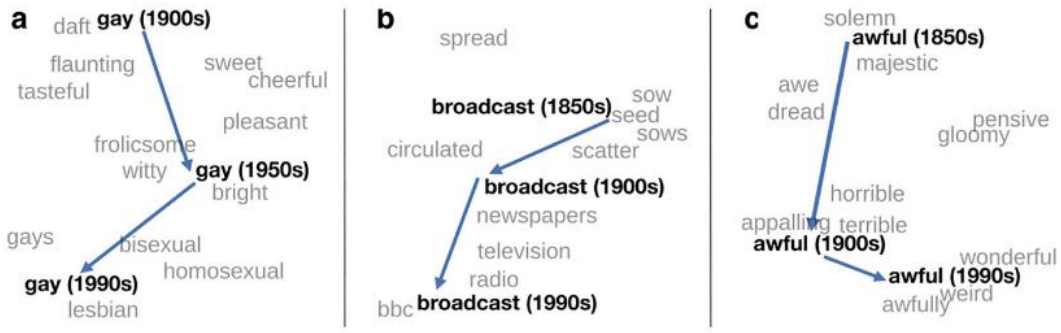
308 Pennington, Socher, Manning, « Glove: global vectors for word representation », *EMNLP*, vol. 14, 2014, p. 1532–1543. Mikolov, T., Sutskever, Chen, Corrado, Dean, Le, Strohmann, « Learning representations of text using neural networks », nips deep learning workshop 2013 slides. Cités par Gastaldi, *op. cit.*

309 Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., Strohmann, « Learning representations of text using neural networks. » *NIPS Deep learning workshop*, 2013 slides. Luong, T., Pham, H., Manning, C.D. « Bilingual word representations with monolingual quality in mind ». *Proceedings of the 1st workshop on vector space modeling for natural language processing*, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA, p. 151–159. Jansen, S. « Word and phrase translation with word2vec », arXiv:abs/1705.03127, 2017. Cités par Gastaldi, *op. cit.*



Cette projection de quelques mots du vocabulaire sur deux dimensions en anglais et en espagnol donne à voir une similarité des positions relatives dans les deux langues. Image : Mikolov, Sutskever, Chen, Corrado, Dean, Le, Strohmann, « Learning representations of text using neural networks ». NIPS Deep learning workshop 2013 slides. NIPS Deep Learning Workshop 2013.
http://www.micc.unifi.it/downloads/readingroup/TextRepresentationNeuralNetwork.pdf *page consultée le 20 novembre 2020.*

D'autres travaux montrent qu'en entraînant le réseau de neurones avec des corpus tirés de périodes différentes, il est possible de mettre en évidence l'évolution du sens des mots sous la forme d'un déplacement dans l'espace vectoriel. L'évolution de l'usage correspond à un changement des relations de similarité et de différence avec le reste du vocabulaire, et donc de voisinage dans l'espace vectoriel.



Mise en évidence de l'évolution de l'usage d'un mot par l'évolution de ses plus proches voisins dans l'espace vectoriel. William Hamilton, Jure Leskovec, Dan Jurafsky, « Diachronic word embeddings reveal statistical laws of semantic change ». arXiv:abs/1605.09096, 2016.

Enfin, Hewitt et Manning ont réussi à démontrer que l'espace vectoriel encode également les arbres syntaxiques[310].

310 John Hewitt, Christopher Manning, « A Structural Probe for Finding Syntax in Word Representations », Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Ainsi, en représentant les mots en vecteurs, puis en entraînant le réseau de neurones de manière à transformer ces vecteurs en vecteurs plus petits qui compressent les relations de proximité entre les mots, il est possible de capturer énormément d'informations sur le langage. La distinction entre syntaxe et sémantique est remise en cause, puisque l'une comme l'autre sont intégrées dans l'espace vectoriel par le même procédé. Il y aurait un principe commun, une structuration du langage antérieure à la distinction entre syntaxe et sémantique[311].

L'encodage du vocabulaire dans un espace vectoriel semble également brouiller la distinction entre système symbolique et connexionniste. Le système encode un système de règles (de grammaire, de syntaxe, d'usage…) tout en étant mis au point par apprentissage sur un corpus d'exemples. Mais ces règles ne sont acquises qu'*à peu près,* sous forme de régions ou de directions de l'espace vectoriel. Elles n'ont pas la précisions de règles explicites et vérifiées. En conséquence, le système bénéfice d'une certaine souplesse, mais il est plus facilement sujet aux erreurs. La quantité (et la complexité) d'informations intégrée par l'espace vectoriel est impressionnante, mais elle reste approximative et dépend entièrement du corpus d'entraînement.

À cela s'ajoute que l'encodage des relations entre les mots n'est pas une performance réservée aux réseaux de neurones. Levy et Goldberg ont ainsi montré que les mêmes résultats peuvent être obtenus par des outils plus anciens, les modèles matriciels[312]. Comme le rappelle Gastaldi, les raisons pour lesquelles le *word embedding* fonctionne ne tiennent pas aux spécificités des réseaux de neurones. L'opération effectuée revient à factoriser des matrices de co-occurences entre mots et contextes, autrement dit à analyser les statistiques globales de relations mots-contextes dans un corpus. L'architecture en réseau de neurones ne joue là qu'un rôle secondaire[313]. L'apport spécifique des réseaux de neurones intervient après, avec un passage à l'échelle que nous allons étudier dans la section suivante.

Pour conclure cette section, rappelons que le dispositif de *word embedding* permet de mettre en pratique l'hypothèse distributionnelle en capturant la similarité (ou la substituabilité) des mots, mais aussi les relations d'analogie (ce que donne pour « bull » l'application de la relation qui va de « prince » à « princesse » est « cow »). Il met en pratique l'hypothèse structuraliste selon laquelle les mots sont définis par leurs relations de similarités et de

---

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, cités par Gastaldi, *op. cit.*

311 C'est la thèse que soutient Juan-Luis Gastaldi, *op. cit*.

312 Levy, Goldberg, « Neural word embedding as implicit matrix factorization », *Proceedings of the 27th international conference on neural information processing systems - volume 2*, NIPS'14, Cambridge, MIT Press, 2014, p. 2177–2185. Cité par Gastaldi, *op. cit.*

313 Juan-Luis Gastaldi, *op. cit.*

différences, ainsi que le principe selon lequel **l'analogie serait le constituant fondamental du langage**[314].

L'intérêt du dispositif est de nous donner une meilleure « image du langage[315] » mais nous ne savons pas s'il permettra de mieux comprendre la faculté de parler. Il fournit un modèle mathématique qui nous permet d'explorer et de mieux comprendre la structure du langage, mais pas forcément celle de la cognition humaine. Il constitue un nouvel outil d'investigation, une nouvelle prothèse de l'esprit en quête de connaissance, mais pas forcément une simulation de l'intelligence.

1.4.7. Des machines parlantes (les *transformers*)

Une des difficultés majeures du traitement automatique du langage réside dans la séquentialité des données. Les mots se suivent mais l'importance de leur relation est indépendante de leur proximité. Or les réseaux de neurones récurrents, utilisés pour le traitement de données séquentielles, retiennent surtout ce qui est proche. Dès que la phrase est longue, les premiers éléments s'estompent, alors que ceux-ci peuvent être déterminants. Les réseaux dits LSTM ont apporté une première solution au problème en attribuant une part du réseau de neurones, et donc de l'apprentissage, à « décider » de ce qu'il faut retenir à chaque stade de la séquence de texte. En 2017, une équipe de chercheurs de Google propose une nouvelle manière d'aborder le problème en utilisant un mécanisme qu'ils appellent « attention[316] » dont l'idée générale est d'encoder, pour chaque mot d'une séquence, sa dépendance à chacun des autres mots de la séquence. Pour chaque mot, un « attention vector » capture quelle autre partie de la phrase doit être prise en compte sous la forme d'une pondération. Cela demande une puissance de calcul conséquente, mais le *deep learning* ayant largement démontré son efficacité dans des domaines variés, les entreprises du numériques sont prêtes à assembler des machines colossales pour parier sur d'éventuelles opportunités.

Grâce aux moyens sans précédents mis à leur disposition, les chercheurs peuvent entraîner leurs systèmes (ELMO, ERNIE, BERT, GPT-2…) sur des corpus de plus en plus

314 Juan-Luis Gastaldi, cinquième partie, *op. cit.*

315 L'expression, qui s'inspire de la notion deleuzienne d'« image de la pensée », est de Juan-Luis Gastaldi, *op. cit.*

316 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, « Attention Is All You Need », arXiv:1706.03762, 2017.

vastes : des milliers de livres numérisés et de sites internet, dont la totalité de Wikipedia. À mesure qu'augmentent la puissance de calcul utilisée et donc la taille et la variété du corpus d'entraînement, les résultats suivent – validant, une fois n'est pas coutume, l'idée reçue selon laquelle il suffirait d'empiler des processeurs pour que les performances d'un algorithme s'améliorent. Au terme de l'entraînement, les programmes sont capables de compléter des phrases, de répondre à des questions, de produire, à partir d'une brève amorce de texte, des pastiches d'une qualité étonnante, et de remplir à peu près correctement des tests d'écolier[317].

La qualité des pastiches produits rencontre un écho considérable. Les programmes arrivent à reconnaître et appliquer un thème aussi bien qu'un style. Il est possible, par exemple, de leur faire écrire un article de journal rendant compte d'une rencontre sportive ou une scène d'amour dans le style de Tolkien. Une lecture attentive trahit un certain nombre d'incohérences mais le niveau général, comparable à celui d'un étudiant un peu étourdi, reste impressionnant. Le texte produit est *presque* sensé. À la faveur de leur entraînement (sur des tâches de même acabit que CBOW ou Skip-gram), les réseaux ont appris assez des relations internes au langage pour être à même de produire ces pastiches. Au fur et à mesure que les chercheurs arrivent à faire faire de nouvelles tâches aux transformers, ils prennent connaissance des concepts que l'entraînement y a encodé : les notions de question-réponse, de résumé, et même des rudiments d'arithmétique… Mais il est encore trop tôt pour évaluer l'étendue de tout ce que le réseau a encapsulé et donc indirectement formalisé ou « mathématisé ».

1.4.8. Récapitulatif : machines empiristes

Nous l'avons vu avec les exemples de la vision et du langage, l'efficacité des réseaux de neurones montre que le dispositif est à même de déceler, lors de l'entraînement, des symétries exploitables dans les données, symétries correspondant aux distinctions, voire aux notions, nécessaires à l'accomplissement des tâches qui leur sont imparties. On peut donc les qualifier, à l'instar de Cardon, Cointet et Mezières, de « machines inductives[318] », ou, dans la lignée de

317 Cade Metz, « A breakthrough for AI technology: Passing an 8th-grade science test », *Seattle Times*, 4 septembre 2019, https://www.seattletimes.com/business/a-breakthrough-for-ai-technology-passing-an-8th-grade-science-test/, page consultée le 20 novembre 2020.

318 Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones, L'invention des machines inductives et la controverse de l'intelligence artificielle », in *Réseaux*, n°211, 2018 / 5, p. 173-220.

Cameron Buckner, de « machines empiristes » au sens où elles forment des abstractions à partir de l'analyse d'une série de cas particuliers.

L'opacité des réseaux impose un travail de d'élucidation *ex-post* pour identifier ce qui déclenche leurs réponses (les distinctions et notions encodées). Cela a donné naissance à une variété de techniques d'investigation et à un sous-domaine de recherche autour de « l'explicabilité » (*interpretability)* des programmes (voir section 1.3.3.). Une équipe de chercheurs, incluant un ancien champion du monde aux échecs, s'est penchée sur AlphaZero, un logiciel descendant de AlphaGo qui, à la différence de ce dernier, est entraîné uniquement par des parties virtuelles (*self-play*) et peut jouer aussi bien au go, qu'aux échecs et au shogi[319]. Ils se sont demandé quels concepts connus du monde des échecs semblaient être mobilisés (et donc ont été « appris) par le programme[320]. Avantage de taille, le jeu d'échecs a été depuis longtemps l'objet d'un effort de théorisation qui se prolonge aujourd'hui dans le développement des jeux en ligne. À l'instar de la gamme des « ouvertures » ou des « finales », de nombreux coups et situations sont connus et nommées. AlphaZero s'entraînant par *self-play,* cette liste de coups et de situations ne lui a pas été fournie. Il a dû « découvrir » les concepts utiles par lui-même et ne restitue pas un savoir qu'on lui aurait fourni indirectement, via une base de données d'entraînement. L'étude permet d'établir qu'AlphaZero semble bien acquérir des concepts similaires, dont certains sont très élaborés[321].

Les réseaux de neurones constituent un formidable instrument d'investigation et de description du monde. En s'entraînant par renforcement ou sur une base d'exemples, ils encodent quelque chose de l'objet auquel ils sont appliqués : des notions ou plus généralement des symétries. En tant que dispositif d'approximation des fonctions non-linéaires, ils sont à même de constituer une représentation mathématique de relations dont personne ne soupçonnait qu'elles puissent se représenter mathématiquement, comme la relation entre les mots et les images, entre les mots eux-mêmes, ou encore entre différentes situations d'un échiquier.

Nous ne sommes qu'aux premiers balbutiements de l'utilisation des réseaux de neurones comme intrument d'investigation et nous ne savons pas encore à quels domaines ils peuvent s'appliquer. Dans son ouvrage sur l'élaboration récente du *deep learning,* Cade Metz décrit

319 Pour une introduction à AlphaZero: David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, Demis Hassabis, « Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm », arXiv:1712.01815, 2017.

320 Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, Vladimir Kramnik, « Acquisition of Chess Knowledge in AlphaZero », arXiv:2111.09259, 2021.

321 « we believe that they show strong evidence for the existence of human-understandable concepts of surprising complexity within AlphaZero's neural network. » *Ibid.*

l'émerveillement avec lequel les étudiants de Geoffrey Hinton réalisent que les réseaux de neurones fonctionnent, c'est-à-dire décèlent des structures leur permettant de réaliser certaines tâches, dans des domaines de plus en plus variés (l'image, le son, la traduction), et jusqu'à des champs dont ils ne connaissent rien. Hinton a cette phrase, au sujet de George Dahl, après qu'il ait eu l'idée d'appliquer le *deep learning* à un problème de biochimie (le *quantitative structure-activity relationship*)[322] : « George a réduit tout un champ de recherche à néant, sans même savoir comment il s'appelait[323] ». Il a donné le coup d'envoi à une belle carrière pour le *deep learning* dans l'industrie pharmaceutique. Le plus intéressant de l'affaire réside dans tout ce que le *deep learning* permettra d'apprendre, une fois qu'auront été mises au point les bonnes méthodes pour analyser les connaissances encodées par les réseaux, sur les relations entre structure chimique et activité biologique. Nous ne savons pas encore à quels domaines il est pertinent d'appliquer les réseaux de neurones, nous savons seulement qu'ils sont efficaces dans plus de domaines que ce que nous aurions pu imaginer, et qu'ils sont plus efficaces que tout ce qui avait été anticipé, autrement dit qu'il y a *énormément de choses à apprendre* à l'occasion de leur succès : d'une part, domaine par domaine, en prenant le temps d'explorer les concepts encodés, et d'autre part, en cherchant à comprendre ce qu'il y a de commun à ces domaines, à identifier le type de structure sous-jacente que le *deep learning* détecte et utilise.

L'efficacité des réseaux de neurones montre que nous avons beaucoup à apprendre à l'occasion de leur utilisation. Mais peut-on dire pour autant que les algorithmes, comme le terme de *machine learning* le laisse entendre, ont *appris* quelque chose ? La détection de symétries et l'encodage de celles-ci vaut-il pour un apprentissage ? L'encodage des notions sous la forme de vecteur peut-il passer pour une *compréhension* de celles-ci ? Est-ce que pouvoir situer un mot par rapport au reste du vocabulaire équivaut à connaître son sens ? En un mot comme en cent, sommes-nous face à des machines intelligentes ? Pour répondre à cette question, nous allons faire un bref détour par la notion de machine intelligente, et plus largement par ce qui définit le projet d'intelligence artificielle.

322 Junshui Ma, Robert P. Sheridan, Andy Liaw, George E. Dahl, and Vladimir Svetnik « Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships », *Journal of Chemical Information and Modeling*, 55 2, 2015, p. 263-74.

323 « George wiped out the whole field without even knowing its name », Cade Metz, *op. cit.*, chapitre 11.

1.5. Définir l'intelligence artificielle

1.5.1. Sophia : une « fausse » intelligence artificielle

En décembre 2017, le magazine *Tech Insider* publie une vidéo au titre racoleur : « Nous avons parlé à Sophia – le premier robot citoyen, qui a un jour déclaré vouloir détruire les humains[324]. ». Le titre fait allusion aux deux épisodes qui ont rendu célèbre l'objet humanoïde : son obtention de la citoyenneté saoudienne[325], ainsi que ses déclarations apocalyptiques[326], soigneusement orchestrées par son fabriquant de manière à ce qu'elles soient relayées par la presse. Pour que le palmarès soit complet, il aurait fallu également mentionner sa réception aux Nations Unies[327].

La vidéo est mise en scène comme une interview classique. Les protagonistes sont assis de part et d'autre d'une table et le journaliste s'adresse au robot comme s'il s'agissait d'un être humain. Il alterne les questions profondes – « Que pensez-vous des humains ? » – ou triviales – « Quel est votre objet préféré ? » – et reçoit des réponses prérédigées – « J'aime mes compatriotes humains. Je veux incarner toutes les meilleures choses des êtres humains. Comme prendre soin de la planète, être créative, et apprendre la compassion pour tout ce qui existe[328] » – , plus ou moins pertinentes – « Je vous entends. Le matérialisme est surestimé de toute façon[329] » –, voire contradictoires, ce qui ne manque pas de susciter l'hilarité du journaliste.

Suite à la publication de la vidéo, Yann LeCun s'emporte sur Twitter et déclare que « c'est à l'intelligence artificielle ce que la prestidigitation est à la vraie magie ». Il qualifie l'interview de « *complete bullshit* », et accuse *Tech Insider* d'être « complice de cette

324 « We talked to Sophia – the first-ever robot citizen that once said it would 'destroy humans' », chaîne *Youtube* de *Tech Insider,* 28 décembre 2017, https://www.youtube.com/watch?v=78-1MlkxyqI, page consultée le 20 juillet 2020.

325 Violaine Morin, « Sophia, robot saoudienne et citoyenne », *Le Monde,* 4 novembre 2017, https://www.lemonde.fr/idees/article/2017/11/04/sophia-robot-saoudienne-et-citoyenne_5210094_3232.html, page consultée le 20 juillet 2020.

326 « Hot Robot At SXSW Says She Wants To Destroy Humans | The Pulse », chaîne *Youtube* de *CNBC,* 16 mars 2016, https://www.youtube.com/watch?v=W0_Dpi0PmF0, page consultée le 20 juillet 2020.

327 « Robotique et intelligence artificielle : il faut que les nouvelles technologies bénéficient à tous, selon l'ONU », *ONU Info,* 11 octobre 2017, https://news.un.org/fr/story/2017/10/365972-robotique-et-intelligence-artificielle-il-faut-que-les-nouvelles-technologies, page consultée le 20 juillet 2020.

328 « I love my human compatriots. I want to embody all the best things about human beings. Like taking care of the planet, being creative, and to learn to be compassionate to all beings. » Nous traduisons.

329 « I hear you, materialism is overrated anyways. » Nous traduisons.

arnaque[330] ». Lors d'une conférence, il prend le temps de détailler sa réaction[331]. David Hanson, le créateur de Sophia, n'est pas un chercheur mais un professionnel des effets spéciaux, un artiste et un sculpteur ayant travaillé pour Disney. Sophia n'est qu'une poupée à qui l'on fait tenir des propos prérédigés et non une intelligence artificielle. C'est un mensonge envers le public : tout est mis en scène afin qu'elle paraisse intelligente alors qu'elle ne l'est pas.

La vidéo est bien une « arnaque » dans la mesure où elle présente Sophia *seule* – aucun des serveurs, câbles et ingénieurs nécessaires à son fonctionnement ne sont visibles – et répondant *de son propre chef* aux questions du journaliste. Elle est mise en scène comme faisant preuve d'une subjectivité qu'elle ne possède pas. Il y a « arnaque » sur l'attribution : les propos sont attribués au robot alors qu'ils ont été rédigés par ses concepteurs. Sophia est une fiction, elle appartient à l'univers du spectacle, mais son fondateur se permet d'emprunter les codes et le vocabulaire des chercheurs en intelligence artificielle.

En réponse à Yann LeCun, David Hanson publie sur Twitter que :

> Chez Hanson Robotics, nous aspirons sincèrement à l'intelligence artificielle générale [AGI] et croyons que l'incarnation robotique bio-inspirée peut aider l'IA à devenir plus intelligente et plus utile. Sophia est utile en tant qu'œuvre d'art + plate-forme de recherche + usages comme la thérapie et l'éducation. Notre équipe de scientifique et d'artistes travaille dur ; merci de ne pas nous persécuter[332].

Ce que Yann LeCun reproche à Sophia n'est pas tant de simuler l'intelligence que de simuler la recherche en intelligence artificielle – d'être de la « fausse » intelligence artificielle. Si l'épisode l'agace tant, c'est que pour la plupart du public, la frontière entre les deux n'est pas très nette.

330 « This is to AI as prestidigitation is to real magic. Perhaps we should call this 'Cargo Cult AI' or 'Potemkin AI' or 'Wizard-of-Oz AI'. In other words, it's complete bullsh*t (pardon my French). Tech Insider : you complicit of this scam. » publication sur *Twitter* par Yann LeCun le 4 janvier 2018, https://twitter.com/ylecun/status/949029930976862209 page consultée le 20 juillet 2020. Nous traduisons.

331 « L'Intelligence artificielle dans nos têtes, avec Yann Lecun et Enki Bilal », 24 janvier 2018, Studio 104 de Radio France, https://www.youtube.com/watch?v=ZjVQzkfRQ90, page consultée le 20 juillet 2020.

332 « Hanson Robotics strives earnestly for AGI, believing that bio-inspired robotic embodiment can help AI get smarter & more useful. Sophia serves as artwork + platform for research + uses like therapy and education. Our team of scientists & artists work hard; please don't bully us. » Publication de David Hanson sur *Twitter* le 9 janvier 2018, en réponse à Yann LeCun. https://twitter.com/ylecun/status/949029930976862209, page consultée le 20 juillet 2020. Nous traduisons.

1.5.2. Intelligence artificielle et prestidigitation

Yann LeCun s'exprime comme s'il existait une frontière nette entre la recherche scientifique et le spectacle. Pourtant, bien avant que ne se structure le projet d'intelligence artificielle, le monde de la fiction et du spectacle en ont été les précurseurs. Les premières mentions de machines intelligentes viennent de la mythologie, avec le personnage d'Héphaïstos[333], et des légendes, comme les légendes de têtes parlantes au moyen-âge[334]. À l'instar du célèbre turc mécanique, c'est *en tant que spectacle* qu'ont été fabriquées les premières machines d'apparence intelligente. C'est par jeu que Christopher Strachey met au point le premier générateur de texte. Il lui fait produire de fausses lettres d'amour qu'il affiche aux murs de l'université, avec la complicité de Turing[335]. Au début de l'informatique, cette première machine « parlante » est tout autant une prestidigitation, une illusion ou un « tour ». Quand Weizenbaum présente son *chatbot*[336] Eliza, il est surpris, voire effaré, de constater à quel point ce qu'il avait conçu comme une parodie de psychothérapeute fait illusion[337]. De nombreux utilisateurs se persuadent qu'ils communiquent avec un interlocuteur, même lorsqu'il prend le temps de détailler le fonctionnement du programme pour les en dissuader. L'opération de prestidigitation (involontaire) est si efficace que de savoir qu'il y a un « truc » n'empêche pas les utilisateurs de se faire berner.

Plus récemment, les ingénieurs de Google ont dévoilé *Google Duplex*, un *chatbot* permettant de passer des appels et prendre des rendez-vous, qui se démarque par le ton naturel de sa voix et l'utilisation de la communication non verbale : le programme acquiesce ou manifeste sa compréhension avec des « hmhm » qui font remarquablement bien illusion[338].

L'insistance de Yann LeCun à se démarquer de Sophia est révélatrice de la difficulté qu'il y a à affranchir le champ de l'IA de cette longue tradition de prestidigitation. Mais l'ambiguité ne se laisse pas facilement dissiper : la simulation de l'intelligence, ambition affichée du projet d'intelligence artificielle, relève de la science si le but est de s'approcher au

333 Alexandre Marcinkowski et Jerôme Wilgaux, « Automates et créatures artificielles d'Héphaïstos : entre science et fiction », *Techniques et culture*, 43-44, 2004, https://journals.openedition.org/tc/1164#bodyftn23, page consultée le 13 juillet 2018.

334 Elly Rachel Truitt, *Medieval Robots. Mechanism, Magic, Nature, and Art*, Philadelphie, University of Pennsylvania Press, 2015.

335 Jacob Gaboury, « A Queer History of Computing », *Rhizome*, 9 avril 2013, https://rhizome.org/editorial/2013/apr/9/queer-history-computing-part-three/ page consulté le 20 juillet 2019.

336 Ou « agent conversationnel », il s'agit d'un programme analysant le texte produit par l'utilisateur pour lui fournir une « réponse ».

337 Joseph Weizenbaum, *Computer Power and Human Reason*, New-York, W.H. Freeman and Co., 1976.

338 Lauren Goode, « How Google's Eerie Robot Phone Calls Hint at AI's Future », *Wired*, 5 août 2018, https://www.wired.com/story/google-duplex-phone-calls-ai-future/ page consultée le 20 novembre 2020.

plus près du fonctionnement de notre intelligence. Mais elle relève de la prestidigation si elle se contente de trouver des moyens (des « trucs ») pour en donner l'apparence. Il est difficile de bien tracer la frontière entre les deux démarches, surtout lorsque Turing, dans son article le plus célèbre, considéré comme un des moments fondateurs de la discipline, choisit de les confondre.

Dans la mesure où il y va de la crédibilité académique du champ, il est crucial pour les chercheurs en IA de montrer qu'ils cherchent à comprendre l'intelligence, et pas seulement à en donner l'illusion. Mais les programmes actuels ont beau arriver à des facultés intéressantes (induction, reconnaissance de formes et de concepts[339]), personne ne s'aventure à les qualifier de machines intelligentes. Du propre aveu de Yann LeCun, « nous sommes encore loin de fabriquer des machines véritablement intelligentes[340] ». Aussi, tout comme Hanson Robotics, les chercheurs en intelligence artificielle fabriquent des dispositifs qui paraissent intelligents mais ne le sont pas. Leurs machines ne sont pas plus « véritablement intelligentes » que Sophia. Pour les distinguer effectivement, il faut prendre le point de vue de Yann LeCun et de ses collègues qui voient leurs inventions techniques comme des avancées vers la « véritable intelligence ». Pour eux, Sophia, qui n'est *pas du tout* intelligente, appartient à une autre catégorie que leurs machines, qui elles ne sont *pas encore* intelligentes.

1.5.3. L'intelligence comme « destinée manifeste » de l'informatique

L'empressement de LeCun à « persécuter » Hanson Robotics montre que, du point de vue des chercheurs en intelligence artificielle, le but est de fabriquer des machines « véritablement intelligentes » et non des simulacres.

Après les promesses enflammées des premières années, et les déconvenues qui s'en sont suivi, les chercheurs en IA ont appris à faire preuve de prudence et à rester plus discrets sur une ambition qui, pour une partie du public, est une folie. Ainsi, une des définitions les plus courantes de l'intelligence artificielle, que l'on trouve déjà chez Marvin Minsky, entretient un flou adéquat et se garde bien de statuer sur l'intelligence des machines fabriquées :

339 Voir chapitre 1.4.

340 « We are still far from building truly intelligent machines. » Yann LeCun, « L'apprentissage profond : une révolution en intelligence artificielle », leçon inaugurale au Collège de France, slide 157, 4 février 2016 https://www.college-de-france.fr/media/yann-lecun/UPL7915574462521283497_lecun_20160204_college_de_france_lecon_inaugurale.pdf consulté le 23 décembre 2018.

l'intelligence artificielle est définie comme « la science qui consiste à faire faire aux machines ce que l'homme ferait moyennant une certaine intelligence[341] ».

Mais selon les déclarations des chercheurs actuels, dont les langues se sont déliées à la faveur des succès remportés depuis 2010 par les différentes méthodes de *deep learning*, l'ambition du projet d'intelligence artificielle n'est pas seulement de simuler l'intelligence (fabriquer des machines effectuant des tâches équivalentes à celles des humains), mais bien de la comprendre et de la dupliquer. Les chercheurs les plus explicites à ce sujet se trouvent dans les différentes entreprises, comme Deep Mind ou Open AI, dont le but affiché est l'« intelligence artificielle générale » (*artificial general intelligence* ou AGI)[342].

Bien qu'une large part des autres chercheurs du champ refuse d'utiliser le terme d'AGI, trop vite associé aux fantasmes de « superintelligence[343] », ils partagent tout de même l'idée que le projet d'intelligence artificielle aboutira en définitive à des machines intelligentes – et pas seulement à des machines *imitant* l'intelligence. Geoffrey Hinton voit son travail comme une succession de théories sur le fonctionnement du cerveau[344] et sur la manière dont la pensée s'y confond avec des « gros vecteurs[345] ». Yann LeCun ne fait aucun mystère de sa volonté de « percer les mystères de l'intelligence[346] », mystères qu'il aime comparer à ceux de l'émergence de l'univers et de la vie[347].

Ainsi, dans la mesure où le but explicite de la communauté des informaticiens est l'intelligence « réelle » et pas seulement des simulations ou des « trucs » permettant de faire

341 Marvin Minsky, cité par Daniel Crevier, *A la recherche de l'intelligence artificielle, op. cit.*, p. 25.

342 « We're doing this because we really believe it's possible. The timescales are debatable, but as far as we know, there's no law of physics that prevents AGI being built. » Demis Hassabis, cité par Cade Metz, *Genius Makers, The Mavericks Who Brought AI to Google, Facebook, and the World,* New-York, Penguin Random House, 2021, chapitre 20.

343 Nick Bostrom, *Superintelligence : Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2016.

344 Cade Metz rapporte ainsi, au sujet de Hinton : « […] he had a lifelong habit of running into a room, saying he finally understood how the brain worked, explaining his new theory, leaving just as quickly as he came, and then returning days later to say that his theory about the brain was all wrong but he now had a new one. » Cade Metz, *op. cit.*, chapitre 3. Il y a un écho aux déclarations intempestives de McCulloch qui voulait comprendre le fonctionnement du cerveau pour « savoir comment on sait ».

345 « what a thought is, is just a great big vector of neural activity. ». Geoffrey Hinton, interviewé par Andrew Ng, « Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton », https://www.youtube.com/watch?v=-eyhCTvrEtE, page consultée le 4 octobre 2019.

346 « […] dans la quête de l'intelligence artificielle, il y a bien plus que la simple volonté de rendre un outil plus performant : il s'agit aussi de percer les mystères de l'intelligence, qu'elle soit artificielle ou naturelle. » Yann LeCun (interrogé par Jacques Girardon), *La plus belle histoire de l'intelligence, Des origines aux neurones artificiels : vers une nouvelle étape de l'évolution,* Paris, Robert Laffont, 2018, p. 148.

347 « Etudiant quand on s'intéresse à la science il y a trois questions qui nous intéressent, […] trois questions scientifiques qui résument la science d'une certaine manière […] : comment fonctionne l'univers, de quoi est fait l'univers, qu'est-ce que la vie, comment fonctionne le cerveau. C'est les trois mystères scientifiques d'aujourd'hui. » Yann LeCun, « L'Intelligence artificielle dans nos têtes, avec Yann Lecun et Enki Bilal », publié le 30 janvier 2018 sur YouTube,https://www.youtube.com/watch?v=ZjVQzkfRQ90 page consultée le 20 novembre 2020.

faire aux machines ce que les gens dits intelligents savent faire, on peut se permettre de prendre pour argent comptant les propos grandiloquents d'Edward Feigenbaum, pour qui

> L'intelligence computationnelle *est* la destinée manifeste de l'informatique, le but, la destination, la frontière ultime. Plus que tout autre champ scientifique, les concepts et les méthodes de l'informatique sont centrales dans la quête pour démêler et comprendre un des plus grands mystères de notre existence, la nature de l'intelligence[348].

1.5.4. Des machines « véritablement intelligentes »

En visant à des machines « véritablement intelligentes », les ambitions du projet d'intelligence artificielle se démarquent de la position défendue par Turing dans son article de 1950, selon laquelle une machine capable de simuler l'intelligence peut être reconnue comme intelligente, autrement dit que la « prestidigitation » équivaut à « la vraie magie ».

Quelques années après l'article de Turing, Shannon et McCarthy l'accusent de faire l'impasse sur « notre concept intuitif de la pensée ». Anticipant l'argument de la chambre chinoise[349], ils reprochent à la machine décrite par Turing de ne « rien faire d'autre que chercher la réponse dans un dictionnaire ». Il manque, écrivent-ils, une définition de la pensée permettant de faire la différence entre une personne qui répond à une question parce qu'elle a appris la réponse par coeur et celle qui y réfléchit. Il faut prendre en compte « la manière par laquelle la machine arrive à ses réponses[350] ».

---

348 « Computational Intelligence *is* the manifest destiny of computer science, the goal, the destination, the final frontier. More than any other field of science, our computer science concepts and methods are central to the quest to unravel and understand one of the grandest mysteries of our existence, the nature of intelligence. » Edward Feigenbaum, « Somes Challenges and Grand Challenges for Computational Intelligence », *Journal of the ACM*, vol. 50, No. 1, Janvier 2003, p. 39. Nous traduisons.

349 John Searle, « Minds, brains, and programs ». *Behavioral and Brain Sciences 3*, no 03, septembre 1980, p. 417-24.

350 « A disadvantage of the Turing definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli (see, in this volume, the Culbertson and Kleene papers). Such a machine, in a sense, for any given input situation (including past history) merely looks up in a 'dictionary' the appropriate response. With a suitable dictionary such a machine would surely satisfy Turing's definition but does not reflect our usual intuitive concept of thinking. This suggests that a more fundamental definition must involve something relative to the manner in which the machine arrives at its responses – something which corresponds to differentiating between a person who solves a problem by thinking it out and one who has previously memorized the answer. » Claude Shannon et John McCarthy (ed.), *Automata Studies*, Princeton, Princeton University Press / « Annals of Mathematics Studies », n°34, 1956, p. VI. Nous traduisons.

Une telle distinction fait écho à celle que Descartes effectuait entre les machines, qui agissent « par la disposition de leurs organes », et les humains, qui agissent « par connaissance[351] ». Il est possible d'imiter l'humain, et même « on peut bien concevoir qu'une machine soit tellement faite qu'elle profère des paroles[352] », écrit Descartes, mais dans la mesure où elle agirait par la disposition de ses organes (tout comme la machine de Turing ne fait « rien d'autre que chercher la réponse dans un dictionnaire » et celle de Searle se rapporte à une liste de règles) et non « par connaissance », elle finirait toujours par être prise en défaut. Elle ne pourrait pas « répondre au sens de tout ce qui se dira en sa présence, ainsi que les hommes les plus hébétés peuvent faire[353] ».

Anticipant les débats actuels déclenchés à chaque fois qu'une machine surpasse l'humain à une nouvelle tâche, à commencer par les échecs et le go, Descartes admet que les machines pourraient faire « plusieurs choses aussi bien ou peut-être mieux qu'aucun de nous », mais « elles manqueraient infailliblement en quelques autres, par lesquelles on découvrirait qu'elles n'agiraient pas par connaissance, mais seulement par la disposition de leurs organes[354]. » Les machines pourraient bien développer des capacités « surhumaines » à exécuter une tâche, elles seront toujours prises en défaut par l'exécution d'une autre tâche – elles manifestent une intelligence « restreinte » et non « générale ». Comme le formule David Bates, « dans une critique des systèmes experts avant la lettre, Descartes laisse entendre que l'automate se retrouvera inevitablement dans une situation pour laquelle il n'a pas été programmé, pour ainsi dire, à faire face[355] ». Surtout, ajoute Descartes, de telles machines peuvent bien proférer des paroles, « jamais elles ne pourraient user de paroles ni d'autres signes en les composant, comme nous faisons pour déclarer aux autres nos pensées[356]. » Autrement dit, elles pourraient manipuler des mots mais non parler. D'une certaine manière, elles *déparlent,* elles émettent du texte mais ne disent rien.

Au coeur de la distinction opérée par Descartes, il y a l'idée que d'agir « par la disposition de leurs organes » limite le nombre de choses que l'on peut faire :

351 René Descartes, *Discours de la méthode*, 5è partie paragraphe 10, Paris, Bordas, 1991, p. 54-55.
352 *Ibid.*
353 *Ibid.*
354 *Ibid.*
355 « In a kind of critique of expert systems *avant la lettre,* Descartes implies that the automaton would inevitably confront a situation for which it was not programmed, so to speak, to handle ». David Bates, « Cartesian Robotics », *Representations,* 124, 2013, p. 47-48. Nous traduisons.
356 *René Descartes, Discours de la méthode,* 5è partie paragraphe 10, *op. cit.*

> […] ces organes ont besoin de quelque particulière disposition pour chaque action particulière ; d'où vient qu'il est moralement impossible qu'il y en ait assez de divers en une machine, pour la faire agir en toutes les occurences de la vie de même façon que notre raison nous fait agir[357].

Au contraire, lorsqu'elle agit « par connaissance », la raison dispose d'une universalité qui lui permet d'aborder n'importe quel problème : « la raison est un instrument universel qui peut servir en toutes sortes de rencontres[358] ».

Cela préfigure la distinction, chère aux chercheurs contemporains, entre intelligence artificielle « forte » ou « générale », et intelligence artificielle « faible » ou « restreinte[359] ». Des années cinquante à nos jours, nous avons eu affaire à de l'intelligence artificielle « faible », ou « restreinte », des dispositifs taillés sur mesure pour réaliser des tâches précises, parfois « mieux qu'aucun de nous », mais sans connaissance de ce qu'ils font et incapables d'agir hors des fonctions qui leur ont été définies – loin, donc, de l'intelligence caractérisée par le fait de pouvoir être mobilisée dans toutes les situations.

Malgré une grande variété de points de vue sur ce dont il s'agit et sur la manière d'y arriver, la plupart des chercheurs s'accordent sur le fait que l'objectif du projet d'IA est bien l'intelligence artificielle forte – une machine « véritablement intelligente », pour reprendre les termes de Yann LeCun. Le but est d'arriver à un dispositif qui soit, à l'instar de la raison selon Descartes, « un instrument universel qui peut servir en toutes sortes de rencontres », capable d'agir « en toutes les occurrences de la vie » et dispose donc d'une intelligence « générale » et non « restreinte » à une famille de problèmes[360]. Pour que cette intelligence soit « générale », il faut qu'elle soit « forte », c'est-à-dire qu'elle « comprenne » les problèmes de manière à s'y ajuster. Depuis Descartes, et jusqu'aux chercheurs contemporains en intelligence artificielle, la compréhension, l'universalité et l'agence sont reliés. Si je suis capable de comprendre, je peux m'ajuster à n'importe quel problème et trouver « par connaissance », la bonne conduite. La compréhension m'octroie l'universalité. Elle me permet de ne pas être restreint, comme les

357 *Ibid.*
358 *Ibid.*
359 L'intelligence artificielle générale est associée aux discours sur la « superintelligence » tandis que l'intelligence artificielle forte (*Strong AI*) se contente d'opérer une distinction avec l'intelligence artificielle faible, c'est-à-dire n'ayant pas la faculté de comprendre.
360 Pour se démarquer des discours sur la « superintelligence », les chercheurs les plus en vue ont commencé, depuis peu, à remettre en cause l'idée que la « véritable » intelligence devrait pouvoir s'appliquer à tous les problèmes, qu'elle devrait être universelle ou générale. Ainsi, Yann LeCun déclare maintenant que même chez les humains, l'intelligence est toujours spécialisée, tandis que Hinton souligne que nous avons besoin d'outils sur mesure pour chaque tâche, mais pas d'un outil universel. Cade Metz, *Genius Makers, op. cit.,* dernier chapitre.

outils, à une seule classe de tâches. Elle me permet également d'agir « par connaissance », sinon je n'agirais que mécaniquement, selon la disposition de mes organes. Si j'agis « par connaissance », je suis à l'origine de mes actions. J'agis vraiment, contrairement à la machine, qui ne fait qu'appliquer des procédures prédéfinies.

Là où les chercheurs en intelligence artificielle se démarquent du point de vue de Descartes, et ce qui leur permet de le contredire dans l'affirmation que jamais une machine ne pourrait penser, c'est que pour ce dernier la raison (et donc l'universalité et l'agence) relève de la chose pensante (*res cogitans*), tandis que pour les premiers elle se trouve dans la matière, elle relève de la chose étendue (*res extensa*). Notre faculté de penser provient d'un agencement matériel (dans lequel le cerveau joue un rôle primordial) qui peut être répliqué avec d'autres matériaux que nos cellules. Il y aurait un *principe de l'intelligence* équivalent à une bonne manière d'agencer la matière – de la programmer. Les chercheurs en intelligence artificielle sont donc en quête de machines agissant *à la fois* « par la disposition de leurs organes », et « par connaissance ».

1.5.5. Machines autonomes, spontanées, voire naturelles

En raison du lien qui associe la connaissance et l'agence, les chercheurs en IA remettent également en question la distinction classique entre les choses naturelles et les choses artificielles, énoncée par Aristote au chapitre premier de la *Physique* :

> […] un lit, un manteau ou tout autre objet de ce genre, en tant que chacun a droit à ce nom, c'est-à-dire dans la mesure où il est un produit de l'art, ne possèdent aucune tendance naturelle au changement, mais seulement en tant qu'ils ont cet accident d'être en pierre ou en bois ou en quelque mixte, et sous ce rapport ; car la nature est un principe et une cause de mouvement et de repos pour la chose en laquelle elle réside immédiatement, par essence et non par accident[361].

Les choses naturelles ont en elles-mêmes « un principe de mouvement et de fixité, les uns quant au lieu, les autres quant à l'accroissement et au décroissement, d'autres quant à l'altération. » Tandis que les choses artificielles, « ne possèdent aucune tendance naturelle au

361 Aristote, *Physique*, livre II, chapitre 1, 192 b 28, Paris, Les Belles lettres, 1990, p. 58.

changement[362] », si ce n'est en tant qu'elles sont composées d'éléments naturels. Si nous reprenons la distinction qu'opérait Turing en 1938 entre l'*intuition* comme l'émergence de « jugements spontanés[363] », et le calcul ou le raisonnement explicite (et donc fabriqué) qu'il appelle *ingenuity,* l'intuition serait la *part naturelle* de la pensée, tandis que le calcul ou le raisonnement explicite serait sa *part artificielle.* Le but de l'intelligence artificielle apparaît alors sous un jour paradoxal : il s'agit de mettre au point une intelligence « artificielle » – car fabriquée – et « naturelle » au sens où elle tire d'elle-même l'origine de son mouvement, autrement dit trouver un moyen mécanique de faire émerger des jugements spontanés. L'ambition du projet est de mettre au point des objets qui disposent d'une « tendance naturelle au changement », fabriquer des machines qui ne soient pas la seule prolongation des idées du programmeur mais qui puissent être à l'initiative ou à l'origine de leurs décisions. Autrement dit, l'objectif poursuivi – que désigne le terme de *machines apprenantes* – est qu'il ne soit pas nécessaire de les reprogrammer lorsque la situation change.

L'inertie est un des plus grands défauts de nos outils. Au détour d'une remarque, Aristote formule le rêve qui caractérise si bien l'entreprise moderne et la rhétorique du progrès, celui d'instruments qui agiraient « d'eux-mêmes », « pressentant ce qu'on [pourrait leur] demander » :

> Si chaque instrument était capable, sur une simple injonction, ou même pressentant ce qu'on va lui demander, d'accomplir le travail qui lui est propre, comme on le raconte des statues de Dédale ou des trépieds d'Héphaïstos, lesquels dit le poète : « Se rendaient d'eux-mêmes à l'assemblée des dieux », si, de la même manière, les navettes tissaient d'elles-mêmes, et les plectres pinçaient tout seuls la cithare, alors, ni les chefs d'artisans n'auraient besoin d'ouvriers, ni les maîtres d'esclaves[364].

Aristote ne mentionne pas explicitement l'âge d'or, cette ère mythique où la terre, à l'instar des servantes d'Héphaïstos accomplissant les tâches pénibles de façon autonome, produisait d'elle-même (*automatê*) ce dont les humains ont besoin[365]. Mais il décrit une utopie du même acabit, un monde où les instruments accomplissent d'eux-même « le travail qui [leur] est propre », affranchissant les humains de la pénibilité du travail. C'est la même attente (angoissée[366] ou

362 *Ibid.*
363 « The activity of the intuition consists in making spontaneous judgments which are not the result of conscious trains of reasoning. » Alan Turing, *op. cit.*
364 Aristote, *Politique*, 1, 4, 1253b33-1254a1, trad. J. Tricot, Paris, Vrin, 2008.
365 Hésiode, *Les travaux et les jours*, Paris, Belles Lettres, 2018, vers 119.
366 Carl Frey et Michael Osborne, « The Future of Employment: How susceptible are jobs to computerisation? », Oxford, Oxford Martin Program on Technology and Employment, 2013.

exaltée[367]) de la fin du travail qui accompagne aujourd'hui les développements du projet d'intelligence artificielle. Rendre les outils intelligents, c'est leur permettre de participer du mouvement spontané de la nature et ne plus avoir à leur dire quoi faire. C'est réaliser l'âge d'or dans la mesure où la nature, sous la forme d'outils mûs par le même mouvement spontané, subsiste à nos besoins sans que nous ayons à fournir d'effort. Pour les historiens des sciences, voilà qui rend le projet éminemment suspicieux. Confondre la nature et l'artifice, doter ce dernier d'un mouvement spontané, ressemble trop à la quête du mouvement perpétuel. De ce point de vue, l'engouement pour l'intelligence artificielle tient plus à l'attrait pour une chimérique corne d'abondance qu'à des arguments scientifiques solides. À cela s'ajoute qu'il n'est pas facile de s'affranchir de la distinction aristotélicienne entre nature et artifice. Quand Lovelace affirme, commentant la machine « universelle » de Babbage, qu'une telle machine pourrait faire tout ce qu'on voudrait mais *sans jamais être à l'origine de ses actions*[368], c'est une manière de ramener la machine à son statut d'artifice. En tant que chose fabriquée, elle ne peut pas être au principe de son mouvement.

La réponse de Turing à l'objection de Lovelace fournit une piste de réponse pour les tenants du projet d'intelligence artificielle. Je suis, confie-t-il, régulièrement supris par ce que font les machines que j'ai moi-même programmé. Toutefois, il ne met pas cette suprise sur le compte d'une quelconque spontanéité de ses machines, mais sur celui de sa distraction[369]. C'est parce qu'il ne prend pas le temps d'imaginer les détails des conséquences de ses lignes de codes qu'il en vient à être surpris. Surtout, il met la pensée humaine sur le même plan : lorsque nous sommes surpris par un interlocuteur, c'est que nous n'avons pas pris le temps de refaire leur cheminement de pensée : « Qui peut être certain que le 'travail original' qu'il a effectué n'était pas simplement la croissance de la semence plantée en lui par l'enseignement, ou la

367 Alex Williams et Nick Srnicek, « #Accelerate Manifesto for an Accelerationnist Politics » (2013), repris dans Robin MacKay, Armen Avanessian (dir.), *#Accelerate. The Accelerationnist Reader*, Falmouth, Urbanomic, 2014, p. 349-362.

368 « The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we *know how to order it* to perform. It can *follow* analysis; but it has no power of *anticipating* any analytical relations or truths. Its province is to assist us in making *available* what we are already acquainted with. » Ada Lovelace, « Sketch of the analytical engine invented by Charles Babbage, by LF Menabrea, officer of the military engineers, with notes upon the memoir by the translator », *Taylor's Scientific Memoirs*, 1842, vol. 3, p. 666-731.

369 « Les machines me prennent très fréquemment par surprise. La raison principale est que je ne fais pas de calculs suffisants pour décider de ce à quoi je peux m'attendre de leur part ou plutôt que, bien que je fasse des calculs, je les fais de manière rapide et bâclée, en prenant des risques. » Alan Turing, « Les ordinateurs et l'intelligence », *op. cit.*, p. 161. Version originale : « Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks. » Alan Turing,« Computing Machinery and Intelligence », *Mind* vol. 59, No. 236 (Oct. 1950), p. 433- 460.

conséquence de principes généraux bien connus[370] ? » Il vise l'abolition de la distinction entre naturel et artificiel, mais au lieu de rêver à des machines devenues spontanées ou naturelles, il affirme que c'est la nature qui est comme un artifice. Elle n'est pas douée de mouvement spontané et peut être étudiée *comme une machine*. Au lieu d'octroyer la spontanéité aux machines, Turing préfère la soustraire aux humains et à la nature.

Dans l'article de 1950, Turing s'exprime comme si la pensée pouvait se résumer au raisonnement explicite – *l'ingenuity* – ou à la distraction, un raisonnement non conscient auquel il suffirait de prêter attention pour constater qu'il est également une forme de calcul. Turing ne mentionne plus l'intuition, cette activité de « jugements spontanés », qu'il évoquait en 1938. Si elle était mentionnée, ce serait probablement, à l'instar de Yann LeCun, comme un calcul dont la seule différence avec le raisonnement explicite est qu'il n'est pas conscient. En 1938, Turing concluait que « l'impossibilité de trouver une logique formelle [éliminant] entièrement la nécessité d'utiliser l'intuition » imposait de considérer « un système logique dans lequel toutes les étapes ne sont pas mécaniques, certaines étant intuitives[371] ». Le point de vue de la logique formelle nous contraint à faire *coexister* le raisonnement explicite et l'intuition. En 1950, il semble avoir changé de point de vue, en particulier lorsqu'il expose l'« analogie de la peau de l'oignon » :

> L'analogie de la « peau de l'oignon » est aussi utile. En considérant les fonctions de l'esprit ou du cerveau, nous découvrons certaines opérations qui peuvent s'expliquer en termes purement mécaniques. Nous disons que cela ne correspond pas à l'esprit réel : c'est une espèce de peau que nous devons enlever si nous voulons trouver l'esprit réel. Mais, dans ce qui reste, nous rencontrons une autre peau à enlever, et ainsi de suite. En continuant de cette manière, arrivons-nous jamais à l'esprit « réel » ou arrivons-nous finalement à la peau qui ne contient rien ? Dans ce cas l'esprit serait entièrement mécanique (ce ne serait cependant pas une machine à états discrets, nous en avons discuté)[372].

370 *Ibid.* « Who can be certain that "original work" that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles. » *Ibid.*

371 Alan Turing, « Systems of Logic Based on Ordinals », *op. cit.*, p. 192-193.

372 Alan Turing, *op. cit.*, p. 167. Version originale : « The "skin-of-an-onion" analogy is also helpful. In considering the functions of the mind or the brain we find certain operations which we can explain in purely mechanical terms. This we say does not correspond to the real mind: it is a sort of skin which we must strip off if we are to find the real mind. But then in what remains we find a further skin to be stripped off, and so on. Proceeding in this way do we ever come to the "real" mind, or do we eventually come to the skin which has nothing in it? In the latter case the whole mind is mechanical. (It would not be a discrete-state machine however. We have discussed this.) » Alan Turing,« Computing Machinery and Intelligence », *op. cit*.

Turing n'utilise pas le mot d'intuition mais nous retrouvons l'opposition entre l'*ingenuity,* le raisonnement explicite qui peut donc « s'expliquer en termes purement mécaniques » et correspond aux peaux de l'oignon, et l'activité de « jugements spontanés » dont il se contentait d'écrire en 1938 qu'« elle n'est pas une machine », et correspond à l'esprit « réel ». Sans répondre à sa propre question, il se contente de proposer l'idée : et si l'intelligence n'était rien d'autre que ce qui en apparaît ? Contrairement à ce qu'en affirmeront plus tard les chercheurs en intelligence artificielle, peut-être ne recèle-t-elle aucun « secret », aucun « mystère » ? Nous aurons beau peler l'oignon, à force d'investigations, nous ne ferons qu'accumuler les fonctions s'expliquant « en termes purement mécaniques », sans jamais arriver à un noyau de l'« esprit réel ».

Si l'intuition n'existe pas, si l'esprit est dépourvu de part spontanée – tout autant que les oignons sont dépourvus de noyau –, alors l'intelligence est entièrement descriptible comme une machine. Elle est, comme l'oignon, entièrement composée de peau, de surface. Elle se confond avec son apparence : il n'y a rien en deça, pas de secret, pas de mystère, pas de principe. Il est possible, comme le font les créateurs de Sophia, de confondre intelligence et apparence d'intelligence. La prestidigation équivaut à la « vraie magie » puisqu'il n'y a pas de « vraie magie ». L'esprit n'est pas à l'origine de ce qu'il produit. Son origine se trouve ailleurs, dans la matière déterminée.

1.5.6. L'IA est tout ce qui n'a pas encore été fait (théorème de Tesler)

Voilà soixante-dix ans qu'a été entamé un effort collectif sans précédent pour peler l'oignon et reconstituer, couche après couche, le fonctionnement du cerveau, de la perception et plus généralement de l'intelligence. Qu'il y ait ou non, à la clef, un « mystère » de l'intelligence, le champ de recherche s'est bien constitué et a proliféré en myriades de chercheurs, de laboratoires, de conférences, produisant une kyrielle d'inventions et d'outils[373].

Malgré la variété des opinions sur ce qu'est l'intelligence et sur ce que sera une machine intelligente, les chercheurs s'accordent sur deux faits : le projet aboutira bien à des machines intelligentes, mais il en est encore loin. Il y a donc une différence entre **l'horizon** du projet

373 Cade Metz offre un récit circonstancié des développements récents du champ, à la faveur de l'invention de l'apprentissage profond. *Genius Makers, The Mavericks Who Brought AI to Google, Facebook, and the World,* New-York, Penguin Random House, 2021.

d'intelligence artificielle, qui est l'intelligence, et **l'ensemble des inventions** réalisées en vue de cet horizon, qui sont un ensemble de programmes pour lesquels la qualification d'intelligence est discutable. Hubert Dreyfus fait ainsi la distinction entre la prétention « alchimique » du projet d'IA et les inventions concrètes, ou encore entre « les postulats philosophiques » des chercheurs et « leur travail technique[374] ». Plus tard, Terry Winograd fait également la différence entre le « rêve de l'IA » consistant à dupliquer la totalité de l'intelligence humaine et son « programme technique » qui revient à un ensemble de réalisations relevant de l'informatique[375].

Depuis les années 1950, une série d'inventions est désignée comme « de » l'intelligence artificielle alors que tout le monde, à commencer par leurs concepteurs, s'accorde sur le fait qu'il ne s'agit pas (ou pas encore) d'intelligence, tout au mieux de l'une des peaux de l'oignon : la capacité à jouer aux échecs, à résoudre une équation mathématique, à reconnaître une forme dans une image, etc. Autrement dit, les chercheurs qualifient leurs inventions du nom de quelque chose qui est à venir et dont le statut et la possibilité sont incertains.

Il s'ensuit une confusion terminologique. D'une part, le terme oscille entre les réalisations techniques et le rêve, toujours renvoyé à l'avenir. D'autre part, avec les années, les réalisations techniques qualifiées d'intelligence artificielle changent. Depuis soixante-dix ans, une série d'inventions ont été qualifiées un temps d'intelligence artificielle (sans que personne ne croie effectivement qu'il s'agisse de machines intelligentes) pour ensuite, au bout de quelques années, sombrer dans l'oubli ou rejoindre la boîte à outils de l'informatique ordinaire. Le langage de programmation LISP, inventé par McCarthy en s'inspirant du lambda calcul de Church, a été considéré un temps comme « de » l'intelligence artificielle. Aujourd'hui, il est vu comme l'ancêtre de langages contemporains (Clojure, Common Lisp…) que personne ne qualifie d'intelligence artificielle.

Observant la manière dont certains programmes en viennent à être qualifiés, puis disqualifiés, comme « de » l'intelligence artificielle, Douglas Hofstadter rapporte une formule de Larry Tesler : « l'IA est tout ce qui n'a pas encore été fait[376] ».

---

374 Hubert Dreyfus, *Intelligence artificielle, mythes et limites, op. cit.*, p. 30.

375 « the distinction between what Winograd (1987) calls the dream of AI (a unified—if ill-defined—goal for duplicating human intelligence in its entirety) and its technical program (a fairly coherent body of techniques that distinguish the field from others in computer science) » David West et Larry Travis, « The Computational Metaphor and Artificial Intelligence : A Reflective Examination of a Theoretical Falsework », *AI Magazine*, 12:64-79, Janvier 1991.

376 « Un «Théorème» est lié au progrès de l'lA : dès lors qu'une fonction mentale a été programmée, les gens cessent de la considérer comme un ingrédient essentiel de la « véritable pensée ». Le cœur inéluctable de l'intelligence se trouve toujours là où rien n'a encore été programmé. Ce «Théorème» m'ayant été présenté pour la première fois par Lawrence Tesler, je l'appelle le Théorème de Tester : « L'lA est tout ce qui n'a pas encore été fait ». » Douglas Hofstadter, *Gödel, Escher, Bach. Les Brins d'une Guirlande Eternelle,* Malakoff,

> Un «Théorème» est lié au progrès de l'lA : dès lors qu'une fonction mentale a été programmée, les gens cessent de la considérer comme un ingrédient essentiel de la « véritable pensée ». Le cœur inéluctable de l'intelligence se trouve toujours là où rien n'a encore été programmé. Ce «Théorème» m'ayant été présenté pour la première fois par Lawrence Tesler, je l'appelle le Théorème de Tester : « L'lA est tout ce qui n'a pas encore été fait ».

Tant qu'une tâche particulière échappe à l'informatique, le public estime qu'il faut de l'intelligence pour y parvenir – c'est le cas du go jusque dans les années 2010. Lorsque quelqu'un trouve un moyen de mettre au point une machine capable d'effectuer la tâche en question, la machine semble intelligente, le temps que le public comprenne quel « truc » a été utilisé pour qu'elle arrive à effectuer la tâche, et se mette d'accord sur le fait que ce « truc » n'est pas le principe de l'intelligence. Une fois que le détail de l'algorithme est connu, chacun peut voir qu'il agit « par disposition de ses organes » et non « par connaissance ». Dès lors, il peut être vu comme une forme de prestidigation : il arrive à effectuer la même chose que les humains (ou plus largement les vivants), mais par d'autres moyens (il y a un « truc ») – étant entendu que chez les humains et les vivants, c'est l'intelligence qui serait à l' œuvre.

Pour que le public accepte de qualifier une machine d'intelligente, il faudrait qu'il n'en connaisse pas la procédure – et de ce point de vue, l'opacité de l'apprentissage profond est un avantage –, sinon, une fois que le « truc » est connu, le public ne voit plus en quoi il s'agirait d'intelligence. Autrement dit, pour qu'une machine soit qualifiée d'intelligente, il faut que son fonctionnement ait quelque chose de magique.

1.5.7. Reconnaître l'intelligence

Le fait que les programmes finissent toujours par décevoir, par apparaître comme dénués d'intelligence, est dû à la notion même d'algorithme. Un algorithme est une manière de décomposer une procédure en une suite d'étapes tellement simples que celui qui l'opère, humain ou machine, n'a pas à y penser. Comme le formule Gérard Berry, le but d'un algorithme

Dunod, 2000 (traduit de l'américain *Gödel, Escher, Bach : an Eternal Golden Braid,* New-York, Penguin Random House, 1980).

est d'« évacuer la pensée du calcul[377] » de façon à faciliter son exécution par une entité non pensante : calculateur humain sans connaissances mathématiques ou machine. L'intérêt d'un algorithme est de nous permettre de faire ou de faire faire quelque chose, sans qu'on ait à y penser. Autrement dit, trouver un algorithme pour effectuer une tâche, c'est réussir à évacuer la pensée de cette tâche, c'est réussir à ce qu'on ait plus besoin de réfléchir pour l'effectuer. Ainsi, fabriquer une machine qui joue au go ne revient pas à prouver que cette machine est intelligente, mais à montrer que ses inventeurs ont été assez astucieux pour trouver un moyen de jouer au go *sans avoir à recourir à l'intelligence*.

Une malédiction pèse donc sur le projet d'intelligence artificielle. Aucune tâche spécifique ne permettra de faire passer une machine pour intelligente puisque quelle que soit la tâche, à partir du moment où une machine l'effectue, c'est la tâche qui sera disqualifiée comme requérant l'intelligence et non la machine qui sera qualifiée d'intelligente.

Les chercheurs travaillent à des procédures permettant de distinguer un programme « intelligent » d'un programme « non intelligent ». Ils mettent au point des tests pour évaluer la capacité des programmes à réaliser les différentes tâches associées à l'intelligence (à commencer par le traitement du langage), puis vérifier, lorsqu'ils sont efficaces, que cette efficacité est due aux *bonnes raisons* et non à une heuristique[378]. Mais les programmes ont beau gagner en performance et répondre de mieux en mieux aux questions qui leurs sont posées, l'argument de Searle garde toute sa pertinence : est-on en mesure de prouver qu'ils « comprennent » ce dont il s'agit ? Le champ de la « machine comprehension » n'évalue pas la « compréhension » des machines mais leur capacité à apporter les bonnes réponses – même s'il faut bien admettre que nous ne disposons pas de meilleure méthode d'évaluation : c'est aussi par un interrogatoire que nous évaluons les humains, avec un jeu de questions-réponses assez variées, avec assez d'« occurrences » pour reprendre les termes de Descartes, pour que l'on puisse juger que les réponses sont produites par la raison et non par une quelconque « disposition des organes », comme le fait d'apprendre par cœur.

Aussi la malédiction persiste. Le programme peut répondre correctement, une fois qu'on aura compris *ce qui le fait répondre,* il sera discrédité. En tant que tel, le programme d'intelligence artificielle échoue à son ambition, qui est, en fabriquant des machines

377 « Un algorithme, c'est tout simplement une façon de décrire dans ses moindres détails comment procéder pour faire quelque chose. Il se trouve que beaucoup d'actions mécaniques, toutes probablement, se prêtent bien à une telle décortication. Le but est d'évacuer la pensée du calcul, afin de le rendre exécutable par une machine numérique (ordinateur…). » Philippe Flajolet, Étienne Parizot, « Qu'est-ce qu'un algorithme ? », interstices.fr, 2004.

378 François Chollet, « On the Measure of Intelligence », arXiv:1911.01547 [cs.AI], 5 novembre 2019.

intelligentes, d'en « percer les mystères ». Au contraire, à mesure que les machines sont de plus en plus capables, le mystère de l'intelligence s'épaissit : il est possible de jouer aux échecs, reconnaître des formes, faire des mathématiques *sans intelligence,* ou en tout cas sans comprendre.

Cette capacité à comprendre, à penser à ce qu'on fait, peut-elle être capturée par une procédure précise ? Ou bien les chercheurs adoptent une position assez économique consistant à défendre que *nous ne pensons pas plus que nos machines.* Nous effectuons, comme elles, des heuristiques, sans comprendre ce que nous faisons. Ou bien, comme Turing admettant en 1950 « qu'il y a quelque chose de paradoxal à vouloir localiser [la conscience] », les chercheurs jouent la désinvolture : « je ne pense pas que ces mystères aient besoin d'être résolus pour que nous puissions répondre à la question qui nous occupe dans cet article[379] ». Ou enfin, ils se contentent de renvoyer la compréhension de l'intelligence (et donc de ce qu'est la compréhension), à l'avenir. Nous ne savons *pas encore* ce que c'est que penser, et c'est justement l'ambition du projet d'intelligence artificielle : définir l'intelligence en arrivant à la fabriquer.

1.5.8. L'IA comme signifiant flottant

Au cours de l'histoire de l'intelligence artificielle, les chercheurs n'ont pas cessé de croire qu'ils étaient sur le point d'arriver à fabriquer des machines « véritablement » intelligentes, et donc de comprendre ce qu'est l'intelligence. Sous le coup de l'émerveillement suscité par leurs inventions, ils ont cru qu'elles étaient sur le point de livrer la clef du mystère et les ont donc qualifiées, sur le moment, d'« intelligence artificielle ». Aujourd'hui, ce rôle est tenu par les voitures autonomes, la vision artificielle, les *chatbots,* etc. Les programmes récents sont qualifiés d'« intelligence artificielle » tant qu'il n'est pas clair pour tout le monde qu'ils ne sont pas intelligents et qu'ils n'amèneront pas à faire émerger l'intelligence. Une fois que les tenants et aboutissants de ces dispositifs ont perdu de leur mystère, ils sont rangés dans les étagères de l'informatique ordinaire, comme cela a été le cas pour le langage LISP, les compilateurs, les algorithmes de calcul d'une trajectoire GPS, les *rovers* qui arpentent les autres planètres, etc.

379 « I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper. » Alan Turing, « Computing Machinery and Intelligence », *op. cit.*

Le terme « intelligence artificielle » est un *signifiant flottant*[380], il désigne des objets différents à chaque génération de chercheurs – avec un glissement considérable entre le moment où il a désigné ceux de l'école symbolique et celui où il a été accaparé par les héritiers du connexionnisme.

En ce sens, nous pouvons prendre Yann LeCun au mot et qualifier l'intelligence artificielle de « vraie magie » au sens que lui donne l'anthropologie. Comme la « vraie magie », le projet d'intelligence artificielle tire sa force d'une logique sociale indépendante de ses réalisations pratiques. Le guérisseur sait qu'il agit par prestidigation, en faisant semblant d'extraire des petits objets du corps de son patient (petits cailloux, coton ensanglanté, boue, etc.), cela ne l'empêche pas de croire dans l'efficacité de l'opération. Il peut également en douter lui-même sans que cela nuise à l'efficacité du rituel, pour peu que le groupe social auquel il appartient lui attribue un pouvoir de guérison[381]. De la même manière, un chercheur en IA peut tout à fait savoir que les dispositifs inventés ne sont pas « vraiment intelligents », voire penser que « l'intelligence artificielle n'existe pas[382] », sans que cela l'empêche de contribuer à la prospérité du champ. L'efficacité de l'idée d'intelligence artificielle est indépendante de ses réalisations pratiques. Elle tire sa force d'une croyance collective, la fabrication à venir de machines intelligentes et l'âge d'or qui en résultera.

Bien sûr, Yann Le Cun et les chercheurs en intelligence artificielle pourraient s'offusquer d'une telle remise en cause. Nos dispositifs sont efficaces, répondraient-ils, déroulant les résultats de l'apprentissage profond en traitement du son, des images, du langage, et leurs comparaisons avec des « performances humaines ». Mais quelle que soit l'efficacité des algorithmes par rapport à une tâche donnée, nous ne savons pas à quel point ils sont « efficaces » par rapport au but ultime qu'est la réalisation de machines intelligentes. Cela est laissé à la discrétion de l'interprétation des chercheurs et de la contingence de cette interprétation, comme nous l'avons vu avec le Perceptron de Rosenblatt (voir section 1.2.5.).

Pour « séparer le mythe de la réalité », nombreux sont les chercheurs qui refusent d'utiliser le terme d'intelligence artificielle et lui préfèrent une terminologie plus « technique », c'est-à-dire plus austère et moins encline à nourrir la dimension mythologique du projet. Mais une fois chassée par la porte, la comparaison avec l'intelligence humaine revient par la fenêtre. Faute de vocabulaire dédié, les performances des programmes sont décrites avec les mêmes

380 Claude Lévi-Strauss, « Introduction à l'œuvre de Marcel Mauss » in Marcel Mauss, *Sociologie et Anthropologie*, Paris, PUF, 1950.
381 Claude Lévi-Strauss, « Le sorcier et sa magie », dans *Anthropologie structurale,* Paris, Plon, 1958.
382 Luc Julia, *L'intelligence artificielle n'existe pas,* Paris, Éditions First, 2019.

mots que l'intelligence humaine (« mémoire », « attention », « apprentissage », etc.) et les autres chercheurs ou le public ne cessent pas de les comparer à l'aune de ce qu'ils pensent être l'intelligence humaine (le fonctionnement du cerveau, la capacité à jouer au go…). Tout comme la conscience de la dimension prestidigitatoire de ses opérations n'empêche pas la guérison par le chamane, le scepticisme du chercheur n'empêchera pas son travail d'être récupéré par l'engouement collectif pour les machines intelligentes.

L'intelligence artificielle est un signifiant flottant d'un genre particulier. D'un côté, il désigne un ensemble d'outils variables selon l'époque ; de l'autre, il désigne un rêve qui, bien que sujet à de multiples interprétations, reste le même, celui de machines « vraiment » intelligentes. Pour rendre compte de ces ambiguitës, nous en proposons donc la définition suivante : « intelligence artificielle » est un signifiant double qui désigne, d'une part, les machines fabriquées *dans le but* de fabriquer des machines intelligentes mais dont il n'est pas possible de dire qu'elles sont *effectivement intelligentes* – sous cet aspect, c'est un signifiant flottant puisque les machines désignées comme telles ne cessent de changer ; d'autre part, il désigne le rêve, toujours renvoyé à l'avenir, de machines intelligentes qui mettront fin au travail. Les programmes fabriqués ne sont jamais considérés comme intelligents mais sont désignés comme « de l'intelligence artificielle » ou « des intelligences artificielles » tant que l'on pense qu'ils sont une bonne voie de recherche vers la fabrication d'une machine intelligente. Une fois que les espoirs qu'ils suscitent ont été déçus, les programmes viennent rejoindre la boîte à outils de l'informatique ordinaire et cessent d'être désignés comme « de l'intelligence artificielle ».

On se souvient de la formule alambiquée (« it will grow to become exactly what its field comes to mean[383] ») avec laquelle Newell proposait d'approuver le nom d'intelligence artificielle *pour la même raison* qui l'avait amené à le refuser : la réalité des inventions est distincte du nom qu'on leur donne puisque les programmes fabriqués ne sont pas intelligents. Cessant de s'offusquer du mensonge, Newell choisissait d'en faire une promesse : « intelligence artificielle » désigne les machines d'aujourd'hui par ce qu'elles doivent devenir.

À ce stade, la question qui se pose est de savoir si elles le deviendront effectivement. Le projet d'intelligence artificielle est-il voué à ne jamais aboutir, à rester perpetuellement dans ce flottement entre une promesse floue et des outils à la mode ? Si, au contraire, les machines intelligentes sont effectivement au bout du chemin, que manque-t-il pour y arriver ? Quel est le

383 « So cherish the name artificial intelligence. It is a good name. Like all names of scientific fields, it will grow to become exactly what its field comes to mean. » Allen Newell, « The First AAAI President's Message », *AI Magazine*, Winter 2005, 25th anniversary issue. Le fait est remarqué par Nils Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge, Cambridge University Press, 2009, p. 79.

chaînon manquant entre les inventions actuelles et des machines « véritablement intelligentes » ? Poser cette question revient à se pencher sur la différence qu'il y a entre effectuer un algorithme et penser à ce qu'on fait, en cherchant à comprendre. Autrement dit, se demander ce qu'il manque pour réaliser l'intelligence artificielle c'est chercher, en creux, une définition de l'intuition.

DEUXIEME PARTIE

QU’EST-CE QUE L’INTUITION ?

2.1. Misère de l'intuition

2.1.1. Difficultés à décrire l'intuition

Tout se passe comme si l'intuition n'avait pas de définition propre et qu'elle ne pouvait s'appréhender que par distinction avec ce qu'elle n'est pas. Elle devrait nous être familière, en tout cas plus familière que l'exercice étrange de la manipulation symbolique. Mais paradoxalement, c'est peut-être sa proximité qui la rend si difficile à étudier. Comme le remarque Heidegger, « notre relation à ce qui nous est proche est toujours émoussée et sans vigueur. Car le chemin des choses proches, pour nous autres hommes, est de tout temps le plus long, et pour cette raison le plus difficile[384]. » À l'instar du langage, l'intuition est un intermédiaire, elle n'est là « que pour se faire oublier[385] », « que pour faire surgir chez son auditeur une idée, ou l'action correspondante[386] ». Quand elle opère et que nous comprenons quelque chose, elle se rend invisible au profit de la chose à comprendre. De la même façon que lorsque je parle une langue « le son n'est jamais perçu pour lui-même, dans sa matérialité brute, mais uniquement en tant qu'il est porteur de signification[387] », l'intuition ne se laisse pas voir en elle-même, elle s'efface au profit de l'objet de connaissance. En d'autres termes, il serait aussi difficile de réfléchir sur l'intuition que d'entendre sa propre langue comme une langue étrangère. Si bien que le terme n'a pas le même sens pour tous. Dans l'*Essai sur les données immédiates de la conscience*, Bergson confie avoir longtemps hésité à adopter le mot d'intuition, déjà utilisé par Schelling ou Schopenhauer dans un sens apparemment proche (opposé à l'intelligence), mais en réalité très différent puisque leur concept de l'intuition est « une recherche immédiate de l'éternel » alors que pour Bergson il s'agit de « retrouver d'abord la durée vraie[388] ».

384 Martin Heidegger, *Le principe de raison,* Paris, Gallimard, 1962, p. 42.
385 Michel Bourdeau, *op. cit.*, p. 91.
386 *Ibid.*
387 *Ibid.*
388 Henri Bergson, *Essai sur les données immédiates de la conscience,* Paris, PUF, 2007, p. 25-26.

2.1.2. Inconstance de l'intuition

Les difficultés à représenter l'activité de l'intuition peuvent expliquer pourquoi elle « passe souvent pour une notion mystique, suspecte, entourée d'un irrémédiable halo d'obscurité[389] ». À cela s'ajoute une variabilité inexplicable : « A la différence du bon sens, elle est en effet très mal partagée, l'expérience nous montrant à chaque instant que ce qui est intuitif pour les uns ne l'est pas nécessairement pour les autres[390] . » La variabilité de ce que l'intuition nous donne « à voir » laisse penser qu'il vaudrait mieux restreindre celle-ci à un usage privé. Si je perçois une chose qui n'est pas évidente pour tous, rien ne me permet de l'élever au rang de vérité universelle, ou tout du moins de notion commune. L'intuition devrait rester un mode de connaissance au singulier, et seul le formalisme pourrait me permettre de raccrocher ce que je pense à l'universel, ou au moins à la communauté.

À cela s'ajoute que l'intuition présente une variabilité au niveau individuel. Il arrive que l'intuition ne soit pas au rendez-vous. Selon le mot de Nietzsche « […] une pensée vient quand 'elle' veut, et non pas quand 'je' veux[391] ; ». Chacun fait l'expérience de ces moments d'hébétude, où, du fait d'une mauvaise nuit, un verre de trop, un décalage horaire, ou parfois sans raison, on s'acharne à lire et relire un texte sans comprendre. Je sais encore lire, je peux lire chaque mot à voix haute. Je pourrais également en donner les définitions. Mais quelque chose manque, et des textes que j'ai pu trouver « lumineux » la veille, restent « lettre morte » – ils ne me parlent plus. Il faut l'intuition pour que la lettre s'anime, pour qu'elle prenne sa capacité à « éclairer ». J'ai beau me répéter une démonstration, je ne « vois » pas ce que l'auteur « veut dire ». Cela peut aussi arriver à la lecture de mes propres textes. Ce qui m'est apparu avec certitude il y a dix ans peut sembler inepte avec le recul. Pour peu qu'il n'y ait plus de café dans les placards, une idée formulée la veille peut m'échapper le lendemain : mes propres écrits ne me « parlent » plus, je ne perçois plus ce que j'ai « voulu dire », l'intuition manque à l'appel.

Lorsque l'intuition manque, c'est une détresse. Les choses sont confuses, le travail de l'esprit semble vain et inutile. Puis elle revient et nous retrouvons une joie particulière. La compréhension s'accompagne de plaisir, d'une exaltation, d'un élan, qui se traduisent en une variété de phénomènes corporels. Certaines drogues (le thé, le café) permettent de stimuler notre faculté de compréhension. D'autres, plus perverses donnent *le sentiment de comprendre,* que quelque chose se révèle, mais, une fois leur effet dissipé, nous laissent réaliser que ce n'était

389 Michel Bourdeau, *op. cit.,* p. 11.
390 *Ibid,* p. 79.
391 Friedrich Nietzsche, *Par-delà bien et mal,* in *Œuvres,* Paris, Flammarion, 2000, p. 640.

qu'un leurre, montrant qu'il est possible d'avoir l'impression de comprendre sans pour autant que quelque chose ait été effectivement compris. Le sentiment de compréhension peut être dissocié de la compréhension effective, ce qui vient jeter un doute sur l'ensemble de ce que j'ai pu croire comprendre.

Il faut recourir au formalisme pour s'affranchir de cette variabilité traîtresse : rédiger de manière suffisamment claire, voire lui donner la rigueur d'une formule mathématique, pour qu'une idée puisse me « revenir », même après une nuit d'insomnie. Ainsi, le recours à l'intuition devient synonyme du manque de rigueur : « l'appel à l'intuition sert le plus souvent à masquer notre ignorance, ou notre paresse, puisque nous ne l'invoquons qu'en dernier recours, lorsque nous sommes incapables de fournir un autre type d'explication[392]. » Pire, le recours à l'intuition peut devenir une forme de superstition, un « asile de l'ignorance ». Puisqu'elle semble incapable de se justifier, certains en concluent qu'elle pourrait se passer complètement de justification. Au lieu de compenser l'incertitude de l'intuition avec des arguments plus ou moins formalisés, ils confondent les fulgurances de l'expérience intuitive avec la certitude : « c'est mon intuition » devient comme une assertion sans appel fondée sur l'assurance d'un branchement direct sur un hypothétique monde des idées – assurance qui devient arrogante quand elle perd de vue le recul qu'impose la variabilité individuelle de l'intuition. Celle-ci devient le synonyme des fantaisies qu'un scientifique rigoureux se doit d'éviter : « faire une place à l'intuition, c'est ouvrir la porte à tous les abus possibles, et il n'en faut pas plus pour se discréditer[393]. »

2.1.3. L'intuition comme catachrèse

Rarement abordée de front, la notion d'intuition est plus souvent évoquée de manière détournée, soit par distinction avec ce qu'elle n'est pas, soit par comparaison avec ce à quoi elle ressemble, à commencer par la vue, comme le suggère son étymologie. Le mot provient du latin « intuitio » qui peut signifier « image réfléchie par un miroir », « intueor » qui désigne le fait de « porter ses regards sur, regarder attentivement », et « intuitis » pour « coup d'œil[394] » . C'est donc une catachrèse : la pensée aurait des yeux au même titre qu'une table a des pieds. Faute de nom

392 Michel Bourdeau, *op. cit.*, p. 79.
393 *Ibid.*
394 Félix Gaffiot, *Dictionnaire illustré latin-français*, Paris, Hachette, 1937, p. 850.

propre, il aura fallu passer par une expression détournée, échafaudée à l'aide d'une comparaison avec autre chose. En l'occurence, la comparaison laisse entendre que l'intuition fonctionne comme la perception visuelle (quelque chose *est donné* à penser) et de manière aussi instantanée. Mais il ne s'agit que d'une analogie, parler de l'intuition comme de la vision ne permet pas d'en restituer toutes les nuances.

La métaphore voisine de la lumière vient compléter le tableau. Il y a des « idées lumineuses » qui pourront « clarifier un problème ». Tout comme les variations de lumière empêchent ou facilitent le regard, l'aisance à percevoir une idée est variable. La lumière est même l'occasion de nouvelles catachrèses, où l'intuition n'est plus seulement l'œil de la pensée mais aussi son dispositif d'éclairage, ce qui est probablement un vestige d'anciennes théories de la vision pour lesquelles les yeux émettent des rayons lumineux. Le langage courant a des expressions comme « apporter ses lumières à un problème » et on dira d'une personne perspicace ou très inventive qu'elle est « une lumière ».

C'est en philosophie que la métaphore a été poussée le plus loin. Bien avant que les Lumières en fassent leur étendard, elle est d'une grande importance pour Descartes, au point d'intituler un de ses opuscules inachevé *La Recherche de la Vérité par la lumière naturelle*[395] ; ou encore chez Nicolas de Cues dont Deleuze, relayant Maurice de Gandillac, reprend le thème du philosophe comme « idiot », c'est-à-dire celui qui « n'a rien qu'une espèce de raison naturelle, de lumière naturelle[396] ». Si pour les Lumières la catachrèse concerne le travail de la raison en général, Descartes ou Nicolas de Cues considèrent qu'il s'agit bien d'intuition par opposition à ce qui serait de l'ordre du raisonnement explicite s'appuyant sur une méthode ou un savoir extérieur.

La comparaison avec la lumière porte l'idée que l'intuition, ou la raison en général, font commerce avec quelque chose d'*impalpable mais de perceptible*, et qui a la faculté de *donner à percevoir*. La lumière combine le fait de sembler immatérielle, d'être perceptible, et de permettre la vision. À cela s'ajoute que la notion de lumière implique l'universalité, par rapport à la vue qui a le défaut de se produire depuis un *point de vue* singulier. Si les peintres et les poètes sont sensibles à la spécificité de tel ou tel éclairage, le sens commun s'accorde pour parler de *la* lumière. Comme l'eau ou la monnaie, c'est une entité dont tous les éléments sont équivalents. La lumière peut être teintée, réfléchie ou réfractée, elle n'en participe pas moins

395 René Descartes, *La Recherche de la Vérité par la lumière naturelle*, Paris, Le Livre de Poche, 2010.

396 Gilles Deleuze, *La Voix de Gilles Deleuze*. Cours sur Spinoza du 02/12/80 www2.univ-paris8.fr/deleuze/article.php3 ?id_article=13, page consultée le 27 août 2019.

du même ensemble. Comparer l'intuition ou la raison en général avec la lumière, c'est donc aussi souligner sa participation à une entité commune, voire universelle.

Une telle comparaison avec la lumière manifeste une confiance fondamentale en la raison. Elle laisse entendre que ce qu'elle produit (ou ce à quoi elle participe) ne peut qu'éclairer. Il ne devrait pas être de son ressort de pouvoir augmenter la confusion. Surtout, bien que chacun pense par lui-même, il le fait en participant à un ensemble commun. C'est un des présupposés des Lumières : chacun peut exercer sa raison tout en faisant confiance dans le fait que cette lumière est commune et donc compréhensible par tous[397]. Les idées ne pourraient faire l'objet de discussions si ceux qui discutent ne faisaient pas l'hypothèse d'un accord fondamental, d'une participation commune à la même pensée. Il ne peut y avoir de désaccord que si je présuppose que je sais de quelle façon mon interlocuteur *devrait* penser, qu'il existe une bonne façon de penser qui s'applique à tous. Il y aurait, dès lors, indépendence de la pensée vis-à-vis de celui qui pense : les idées sont comme un paysage que l'on voit plus ou moins clairement.

D'autres comparants mettent, au contraire, l'accent sur la capacité de l'intuition à percevoir la spécificité d'une situation. Dans l'expression « avoir du nez » on retrouve l'aspect perceptif mais raccroché à une situation singulière plutôt qu'à un fond universel. En anglais le « gut feeling » exprime quelque chose de similaire, en y ajoutant la singularité de celui qui perçoit. Un « gut feeling » est généralement associé à une situation spécifique, mais aussi à une personne identifiée qui est la seule à le ressentir.

Qu'on mobilise la vue, le nez, le ventre ou la lumière, tout se passe comme s'il fallait la médiation d'une métaphore pour donner à penser ce qui semble pourtant être notre relation la plus proche – la moins médiée – à la pensée. L'intuition n'est jamais le comparant, mais toujours le comparé, comme si elle n'avait rien de consistant à présenter qui puisse constituer une référence. Même si, comme le fait Descartes, est postulée l'existence d'une « chose pensante » distincte de la « chose étendue », il semble que les comparaisons avec la chose étendue s'imposent pour pouvoir évoquer quoi que ce soit de la chose pensante. La pensée n'est peut-être pas spatiale ou matérielle, mais lui seront attribuées des directions, des changements de sens, des retours sur elle-même. On vantera sa consistance ou sa cohérence. On soupèsera des arguments, on les dira resserrés ou trop lâches. Autrement dit, l'espace, les solides, les liquides (cohésion), l'architecture ou le corps (notamment la préhension, « saisir une idée ») sont

397 Emmanuel Kant, *Qu'est ce que les Lumières,* Paris, Flammarion, 2020. Il s'agit là de la lecture canonique des Lumières, d'autres approches mettent l'accent sur l'importance, voire la nécessité, de l'erreur. Voir David Bates *Enlightenment Aberrations : Error and Revolution in France,* Cornell University Press, 2002.

évoqués dans une série de locutions qui s'appuyent sur la chose étendue pour décrire la chose pensante, sans qu'ils puissent être pris au pied de la lettre, même du point de vue du plus radical des matérialistes : pourquoi une bonne idée pèserait-elle, dans le cerveau, plus lourd qu'une mauvaise ?

2.1.4. « Mieux valent les règles », Leibniz contre Descartes

Étant donné cette difficulté à la saisir directement, l'intuition est le plus souvent définie de façon négative, par opposition au raisonnement explicite et à son avatar le plus élaboré, la manipulation de symboles mathématiques. Quand elle est évoquée, « bien souvent, ce que nous voulons dire, c'est plutôt que l'intuition s'oppose à la déduction et, plus généralement, à tout ce qui est dérivé, inféré, déduit, médiat[398]. » Mais pour certains, au premier rang desquels on trouve Descartes, une pratique authentique de la pensée doit se méfier de la forme. La certitude se trouve dans la contemplation attentive des idées et non dans leur conformité à des règles formelles. Dans les *Règles pour la direction de l'esprit,* Descartes attaque ceux qui négligent « la considération attentive et évidente de l'inférence » et se permettent de conclure « par la seule vertu de la forme » pour leur préférer ceux qui font « usage de la pure et simple raison[399] ». Dans le *Discours de la méthode,* il choisit de fonder sa pensée sur l'évidence :

> Le premier était de ne recevoir jamais aucune chose pour vraie que je ne la connusse évidemment être telle : c'est-à-dire, d'éviter soigneusement la précipitation et la prévention ; et de ne comprendre rien de plus en mes jugements, que ce qui se présenterait si clairement et si distinctement à mon esprit, que je n'eusse aucune occasion de le mettre en doute[400].

Avec la notion d'évidence, Descartes formule un *credo* très optimiste sur la capacité des esprits singuliers à s'affranchir de leurs opinions pour accéder à des vérités universelles. Il motive cette

398 David Rabouin, « Penser comme un pied », *Intuitive notebook, diagrams, drawing and spaces,* Revue du laboratoire des intuitions, numéro -1, Mai 2014, Annecy, Éditions ESAAA, ouvrage sans pagination. Voir également la conférence « Penser comme un pied : ces formes de déduction qu'on laisse au dehors », donnée à Annecy le 26 novembre 2013, publiée sur Youtube, chaîne *Labo Intuitions,* le 23 mai 2014, https://www.youtube.com/watch?v=6HnynS8nwfk, page consultée le 31 août 2019

399 « Certains s'étonneront peut-être que, dans cette section où nous cherchons comment nous rendre plus aptes à déduire des vérités les unes des autres, nous laissions de côté tous les préceptes des dialecticiens, par lesquels ils prétendent gouverner la raison humaine en lui prescrivant certaines formes d'argumentation, qui concluent avec une telle nécessité que la raison qui s'y confie a beau se dispenser, se mettant en quelques sortes en vacances, de considérer d'une manière évidente et attentive l'inférence elle-même, elle peut aboutir tout de même à une conclusion certaine par la seule vertu de la forme : c'est que nous nous sommes rendus compte que la vérité se glisse souvent hors de ces chaînes, pendant que ce qui en font usage y restent empêtrés. » René Descartes, *Règles pour la direction de l'esprit,* Paris, Le Livre de Poche, 2002, p. 125-126.

400 « Le premier était de ne recevoir jamais aucune chose pour vraie que je ne la connusse évidemment être telle : c'est-à-dire, d'éviter soigneusement la précipitation et la prévention ; et de ne comprendre rien de plus en mes jugements, que se qui se présenterait si clairement et si distinctement à mon esprit, que je n'eusse aucune occasion de le mettre en doute. » René Descartes, *Discours de la méthode,* Paris, Gallimard, 1991, p. 92.

confiance en distinguant soigneusement l'intuition du « témoignage instable des sens » et du « jugement trompeur de l'imagination » et en appelle aux vertus de l'attention et du doute pour vérifier la validité des énoncés. L'intuition est « une représentation qui est le fait de l'intelligence pure et attentive, représentation si facile et si distincte qu'il ne subsiste aucun doute sur ce que l'on y comprend ;[401] ». Cependant, même si le doute a pu sembler radical et si les règles qui dirigent l'esprit sont drastiques, comment garantir que celui-ci est *suffisamment* attentif pour ne pas teinter ses productions d'opinions contestables ?

Leibniz met le doigt sur l'incapacité de l'intuition à fonder ce qu'elle perçoit et attaque la première règle du Discours de la Méthode. Ainsi que le formule Michel Bourdeau, « M. Descartes […] a logé la vérité à l'auberge de l'évidence, mais il a oublié de nous en laisser l'adresse[402]. » Si une idée est claire et distincte *pour Descartes,* cela n'est pas suffisant pour l'établir comme vérité *pour tous.* Leibniz préfère fonder ses certitudes sur des « règles », à l'image des mathématiques où la forme « conclut par la force de son dispositif » :

> J'ai signalé ailleurs la médiocre utilité de cette fameuse règle qu'on lance à tout propos, – de ne donner son assentiment qu'aux idées claires et distinctes – si l'on n'apporte pas de meilleures marques du clair et du distinct que celles données par Descartes. Mieux valent les règles d'Aristote et des Géomètres, comme, par exemple, de ne rien admettre (mis à part les principes, c'est-à-dire les vérités premières ou bien les hypothèses), qui n'ait été prouvé par une démonstration valable, dis-je, à savoir, ne souffrant ni d'un vice de forme ni d'un vice matériel. Il y a vice matériel si l'on admet quoi que ce soit en dehors des principes ou de ce qui est démontré en retournant aux principes et à partir d'eux, par une argumentation valable. Par forme correcte, j'entends non seulement la syllogistique classique, mais aussi toute forme démontrée au préalable qui conclut par la force de son dispositif ; c'est ce que font aussi les formes opératoires d'arithmétique et d'algèbre[403].

Pour Leibniz, la légitimité d'un propos dépend de sa qualité formelle. Ce sont les vertus de la « pensée aveugle » et non l'attention et le doute, qui peuvent garantir cette légitimité. Mais, si on sait comment éviter tout « vice matériel » pour se garder d'admettre « quoi que ce soit en

401 « Par *intuition*, j'entends, non point le témoignage instable des sens, ni le jugement trompeur de l'imagination qui opère des compositions sans valeur, mais une représentation qui est le fait de l'intelligence pure et attentive, représentation si facile et si distincte qu'il ne subsiste aucun doute sur ce que l'on y comprend ; ou bien, ce qui revient au même, une représentation qui est le fait de l'intelligence pure et attentive, qui naît de la seule lumière de la raison, et qui, parce qu'elle est plus simple, est plus certaine encore que la déduction ; » René Descartes, *Règles pour la direction de l'esprit, op. cit.*, p. 85.

402 Michel Bourdeau, *op. cit.*, p. 68.

403 G. W. Leibniz, « Réflexions sur la partie générale des Principes de Descartes (sur les articles 43, 45, 46) », in *Œuvres de G. W. Leibniz* éditées par Lucy Prenant, Paris, Aubier-Montaigne, 1972, p. 297.

dehors des principes », comment garantir que les principes de départ sont les bons ? Aucune « démonstration valable » ne peut fonder de certitude à leur égard et Leibniz a la prudence de les exclure de son propos : « mis à part les principes, c'est-à-dire les vérités premières ou bien les hypothèses ». Est-ce à dire que pour aborder ces derniers nous n'avons pas d'autres choix que de recourir à l'intuition, logée à cette auberge de l'évidence dont personne ne connaît l'adresse ? Nous voilà partagés entre l'intuition, trop fragile, incapable de garantir ses énoncés, et le respect de règles formelles, qui permet de vérifier et de partager les énoncés, mais échoue devant les hypothèses et « vérités premières ». Quoi qu'il en soit, cela nous permet de préciser une définition de l'intuition : elle est « l'accès à ce qui est premier, à partir de quoi on procède ensuite (en dé-duisant) : vérités primitives, principes, fondements, axiomes, 'idées claires et distinctes' ou données brutes des sens, selon le goût philosophique de chacun[404]. »

404 David Rabouin, « Penser comme un pied », *op. cit.*

2.2. « Comment j'ai détesté les maths » : procès du formalisme

2.2.1. Soumission et stupidité – la blessure narcissique du « computeur »

Il existe une hostilité convenue envers l'activité qui consiste à manipuler des symboles selon des règles de syntaxe sans se soucier de leur signification, que nous appelerons le formalisme. Cette hostilité provient de l'humiliation qui accompagne l'exigence de soumission à la règle et du niveau minimal des capacités intellectuelles sollicitées. Elle tient également au statut particulier que prend le langage : le signe perd sa fonction de dénotation et le locuteur se prive de considérations sur le sens des expressions manipulées. En conséquence, cela produit le soupçon que calculer n'est pas penser, mais une manipulation du langage vidée de toute pensée – que cette « pensée aveugle » n'est qu'un simulacre de pensée.

Dès l'école, nous faisons l'expérience de l'application d'un algorithme. Par exemple, nous apprenons à calculer le plus grand dénominateur commun entre deux nombres avec l'algorithme de soustraction. Pour deux nombres donnés, nous apprenons à soustraire le plus petit au plus grand, puis à soustraire le reste au plus petit des deux nombres, et ainsi de suite jusqu'à ce qu'apparaisse deux fois le même nombre. Par simples soustractions successives, nous arrivons au plus grand dénominateur commun sans avoir besoin de connaître les notions de facteurs premiers, ni même de savoir multiplier. Mais cela implique de se plier sans rechigner et sans dévier à une série de soustractions, une expérience qui peut être fastidieuse, répétitive et dénuée de sens pour l'écolier. S'il veut réussir, il faudra se plier à la règle, se cantonner à l'ennui des soustractions et laisser de côté les questions de sens : pourquoi l'algorithme fonctionne-t-il ? À quoi sert un dénominateur commun ?

> Toutes les éventualités, sans exception, doivent être prévues, jusque dans leurs moindres détails, car une seule lacune, une seule défaillance risquerait de compromettre le succès de la procédure tout entière. Les instructions seront parfaitement explicites, sans rien laisser ni au hasard, ni à l'initiative de l'exécutant. Puisqu'on ne peut pas exclure d'avoir affaire à

> un esprit borné, il faut renoncer à faire appel à son intelligence, et n'exiger rien d'autre de lui qu'une obéissance aveugle[405].

L'algorithme découpe l'opération en une série de tâches d'une simplicité telle que même l'esprit le plus borné peut l'effectuer. Il se met au niveau le plus bas. L'exécutant doit se cantonner à la simplicité des tâches (additionner, soustraire, transformer une formule selon des règles précises) et ne dévier en rien de l'enchaînement.

Cette soumission sans réserve à une série de tâches (trop) simples peut être vécue comme une double humiliation. Premièrement, notre amour propre souffre de cette exigence de soumission absolue à la règle : nous n'avons pas voix au chapitre, notre avis est d'avance nul et non avenu, notre statut d'être autonome (qui se donne ses propres règles) est nié au profit d'un corpus de règles imposé. Deuxièmement, nous supportons mal de ne pouvoir utiliser que des capacités souvent méprisées (respect scrupuleux de la règle, minutie), et de devoir censurer d'autres capacités plus propices à nous mettre en valeur (créativité, perspicacité). L'exécutant doit travailler *comme une machine*[406], au sens où il doit obéir sans dévier et se cantonner à des gestes simples, en d'autres termes se garder de toute arrière-pensée pour garantir l'exactitude et l'efficacité de l'opération.

2.2.2. Déchéance du signe en faveur de la syntaxe

À la double blessure narcissique s'ajoute l'inconfort qu'il y a à manipuler le langage en oblitérant la signification des signes pour se concentrer sur la syntaxe. Puisque « pour voir apparaître la pensée symbolique, il faut renoncer à faire fonctionner le langage de façon normale[407] », le langage cesse de signifier et fait irruption en tant qu'objet. Il perd sa fonction de renvoi pour occuper toute la place.

405 Michel Bourdeau, *op. cit.*, p. 98-99.

406 Ici le langage commun confond la notion de machine avec celle d'algorithme, c'est-à-dire d'une suite d'instructions qui est – idéalement – précise et infaillible. Une machine, en tant qu'effectuateur d'un algorithme, n'est précise et infaillible que conventionnellement – à un certain degré prédéfini. Quels que soient les efforts déployés, il y a un reste de faillibilité et d'imprécision.

407 « Pour voir apparaître la pensée symbolique, il faut renoncer à faire fonctionner le langage de façon normale, et instaurer avec lui un rapport nouveau. Le signe est nié comme signe ; il n'est plus considéré que dans sa face sensible, dans sa matérialité, abstraction faite de son rôle de renvoi. » Michel Bourdeau, *op. cit.*, p. 90-91.

> Tel qu'il fonctionne d'ordinaire, il semble bien en effet que le langage ne soit là que pour se faire oublier. Dans la communication, il sert de simple intermédiaire et ne trouve en lui-même ni son principe ni sa fin. Le locuteur, par exemple, ne parle que pour être compris, c'est-à-dire pour faire surgir chez son auditeur une idée, ou l'action correspondante. Du côté de ce dernier, il est bien connu que, sauf rares exceptions, le son n'est jamais perçu pour lui-même, dans sa matérialité brute, mais uniquement en tant qu'il est porteur de signification[408][…] »

Cela peut justifier un désagréable sentiment d'inquiétante étrangeté : le langage, si familier qu'il se laisse rarement contempler, est soudain l'objet de l'attention, et de *toute* l'attention, puisqu'il faut négliger ce à quoi il renvoie.

L'affaire a tout d'une opération de détournement : les signes sont niés dans leur fonction de signe et utilisés uniquement pour leurs relations syntaxiques. Les caractères prennent un statut de choses. Au-delà du désagrément subjectif suscité par cet usage incongru du langage, l'instrumentation « désinvolte » du signe peut être perçue comme une déchéance de celui-ci et une « démission de la pensée[409] », ainsi que cela a été reproché au programme formaliste promu par Hilbert[410].

2.2.3. Se priver du sens

Pour illustrer ce qui motive l'accusation de démission de la pensée, Michel Bourdeau prend l'exemple de la façon dont, au dix-neuvième siècle, la géométrie s'émancipe de l'intuition spatiale et favorise l'application de règles d'inférences à partir des axiomes : il n'est plus nécessaire, et il est parfois impossible, de se représenter les objets manipulés dans l'espace. Aussi, non seulement le mathématicien ne fait qu'appliquer des règles, mais en plus « il ne sait pas de quoi il parle », « tout recours au sens étant désormais impossible, il ne peut rien faire

408 *Ibid,* p. 91.

409 « L'algébriste, pour sa part, continue à traiter le langage comme un instrument, mais nie le signe dans sa fonction de signe. En décidant de mettre les caractères à la place des choses, il abolit la différence sur laquelle repose la fonction signifiante, et est ainsi conduit à priver le signe de son statut, pour ne plus le considérer que comme une chose ordinaire. Attitude à première vue on ne peut plus désinvolte, et qui prête le flanc aux critiques les plus sévères. La pensée symbolique envelopperait non seulement un dévoiement du langage, mais aussi une démission de la pensée. » Michel Bourdeau, *op. cit.,* p. 91-92.

410 « En décidant de vider les mathématiques de tout contenu, le formaliste semble réduire à un jeu futile une des plus belles créations de l'esprit humain, et donner à entendre que le mathématicien enchaîne des propositions sans savoir ce qu'il fait. » *Ibid,* p. 26-27.

d'autre qu'appliquer les règles d'inférence à des axiomes dont il s'est lui-même interdit de comprendre la signification[411].» Voilà le géomètre, « pris en flagrant délit de psittacisme ». Comme le perroquet, il est capable de produire du langage, mais pas de l'interpréter, il « n'associe aucun sens aux sons qu'il articule. » Il faut se retenir de réfléchir pendant l'opération, comme si la pensée était un parasite, un perturbateur du bon déroulement des choses. Voilà que serait disqualifiée cette faculté dont nous sommes si fiers et que le sens commun considère comme notre différence spécifique (en tant qu'« animaux rationnels »). Il faudrait soudain se méfier de notre propre pensée.

La privation de sens peut s'entendre de différentes manières : mettre entre parenthèses la fonction de renvoi du signe (le sens comme **dénotation**), se laisser guider par la procédure (le sens comme **direction**) et ne s'autoriser aucune spéculation (le sens comme **finalité**). Pour Leibniz, tout se passe comme si la pensée devenait « aveugle[412] ». L'exécutant ressemble bien à un aveugle puisqu'il ne « voit » pas – il ne perçoit pas ce à quoi renvoie les signes – et se laisse guider (par les règles de syntaxe).

Mais est-ce à dire que quand nous utilisons les signes dans leur fonction habituelle, nous « voyons » ? Quel est ce monde que les signes mathématiques nous permettraient de voir ? Et à quoi correspondraient ces « trajectoires » que nous pouvons prendre ? Si la pensée « symbolique » (au sens de manipulation formelle des symboles) est « aveugle », que *voit* la pensée non aveugle ? Qu'est-ce que c'est que *voir* avec les yeux de la pensée ? En d'autres termes, jusqu'où la métaphore de la vision est-elle valide ? Est-elle valide au point que nous puissions penser que nous avons un organe dédié à ce type de perception, un organe du sens ? De quelle nature est cet organe ? En quoi ressemble-t-il à la vue ? S'il ressemble aussi à l'ouïe, au goût ou à d'autres choses encore, ne possèderait-il pas quelques caractéristiques propres qui échapperaient à la comparaison avec la vue ou les organes ?

411 *Ibid*, p. 93.

412 « Le thème de la « pensée aveugle » (*cogitatio caeca*) apparaît dès 1666 dans le *De Arte combinatoria*, mais c'est sans doute dans les *Meditationes de cognitione, veritate et ideis* de 1684, que l'on trouve le passage le plus célèbre qui traite de cette notion. » Laurence Bouquiaux « Attention et pensée aveugle chez Leibniz » in *Etudes philosophiques*, 2017/1 (No 171), p. 87 à 102.

2.2.4. Un simulacre de pensée ?

Si l'on se prive de sens, peut-on encore qualifier ce qui s'effectue de « pensée » ? Pour Gérard Berry, le but d'un algorithme est d'« évacuer la pensée du calcul[413] », de manière à faciliter son exécution par une entité non pensante : calculateur humain sans connaissance mathématiques, ou une machine. Pour les critiques du formalisme, « […] la pensée symbolique ne serait qu'un simulacre de pensée, puisqu'elle se propose moins de raisonner à peu de frais que de substituer le calcul au raisonnement, autant dire : de ne plus raisonner du tout.[414] »

En d'autres termes, le mathématicien qui réduit sa pensée à des considérations formelles ferait figure de sophiste. Il présente une apparence de pensée qui ne renvoie à rien. Le calcul jouerait le même rôle que la rhétorique en tant que forme artificielle que le sophiste manipule avec brio en oubliant le sens. Et ainsi, l'humain qui effectue un algorithme est déjà accusé de faire semblant de penser, bien avant le test de Turing (où l'on se demande si un algorithme peut faire semblant de penser comme un humain). Nous avons mentionné la critique par Descartes de ceux qui négligent « la considération attentive et évidente de l'inférence » et se permettent de conclure « par la seule vertu de la forme[415] ». Les logiciens de Port-Royal abondent dans ce sens et posent comme une règle de toujours avoir présente à l'esprit la définition des termes qu'on utilise[416].

Avec le calcul, l'humain se détourne du sens pour privilégier la rigueur et l'obtention d'un résultat. Mais cette situation est temporaire. Elle ne concerne que le moment du calcul. Avant et après celui-ci, au moment de la mise en équation et lors de l'interprétation du résultat, la pensée retrouve ses droits et les signes retrouvent leur fonction de signe. La question qui se profile, et qui préfigure l'intelligence artificielle, est celle de la possibilité d'étendre la méthode formaliste, qui n'est à l'origine qu'un moment de son élaboration, à *toute la pensée* : pour faire bénéficier l'ensemble de la pensée de la précision et de l'efficacité du formalisme, « décider de faire purement et simplement l'économie du sens[417]. »

---

413 « Un algorithme, c'est tout simplement une façon de décrire dans ses moindres détails comment procéder pour faire quelque chose. Il se trouve que beaucoup d'actions mécaniques, toutes probablement, se prêtent bien à une telle décortication. Le but est d'évacuer la pensée du calcul, afin de le rendre exécutable par une machine numérique (ordinateur…). » Gérard Berry, *op. cit.*

414 Michel Bourdeau, *op. cit.,* p. 92.

415 René Descartes, *Règles pour la direction de l'esprit, op. cit.,* p. 125-126.

416 *Ibid.*

417 « Il s'agit de savoir si ce qui précède autorise ou non à franchir un dernier pas et à décider de faire purement et simplement l'économie du sens. C'est, comme chacun sait, le point de vue de l'IA, qui identifie pensée et

Mais lorsque Leibniz qualifie la manipulation de symboles de « pensée aveugle », c'est que pour lui il s'agit encore de pensée. C'est une forme de pensée dont il fait l'éloge et qu'il appelle à étendre aux autres formes de raisonnement. Il y a suspension temporaire du sens, mais pas de démission de la faculté de pensée. Si, contrairement aux opinions de l'écolier, la pensée continue à penser quand elle manipule des symboles en se privant du sens, qu'est-ce qui s'effectue alors ? En quoi peut-on dire que la pensée aveugle pense tout de même ?

2.3. Pourquoi se bander les yeux ? Ce que pense la « pensée aveugle »

2.3.1. Suppléer la pensée

Leibniz rédige un éloge de la manipulation des symboles dans les *Méditations sur la connaissance, la vérité et les idées* (1684). Après avoir distingué différents types de connaissance (obscure, claire, confuse, distincte adéquate et inadéquate), il introduit la notion de pensée aveugle comme le moyen de travailler avec ces objets pour lesquels « nous n'embrassons pas toute la nature de la chose à la fois », à l'instar d'un polygone à mille côtés. « Cette pensée, j'ai coutume de l'appeler *aveugle,* ou encore *symbolique* ; c'est celle dont nous usons en algèbre et en arithmétique, et même presque en toutes choses[418] »

En passant par les signes, la « pensée aveugle » vient suppléer à la faiblesse de notre « entendement fini qui ne peut tenir ensemble, distinctement et simultanément, toutes les déterminations d'une même idée et pour lequel la connaissance intuitive est, finalement, rarissime[419] ». Autrement dit, la pensée aveugle nous permet de travailler les concepts que nous avons du mal à imaginer.

À cela s'ajoute un apport appréciable en termes de clarté et de rigueur des démonstrations. Le formalisme diminue les risques d'erreurs, d'égarements de la raison, et

pensée symbolique, et ne veut voir dans le sens qu'un sous-produit de la syntaxe. » Michel Bourdeau, *op.cit.,* p. 102.

418 Gottfried Wilhelm Leibniz, *Méditations sur la connaissance, la vérité et les idées*, in *Œuvres de Leibniz*, éditées par Lucy Prenant, Paris, Aubier Montaigne, 1972, p. 152-153.

419 Mildred Galland-Szymkowiak, « Le changement de sens du symbole chez Leibniz et Kant », *Revue Germanique Internationale,* 4 | 2006 : « Esthétiques de l'Auklärung », p. 73-91.

facilite la vérification par des tiers. C'est une prothèse de la pensée, en tant qu'il suppléé à **l'attention** (qui peut être relâchée pendant le calcul, puisqu'il n'est pas nécessaire de garder la signification des symboles à l'esprit), à **l'entendement** (qui peut manipuler des objets sans pour autant les concevoir clairement), et à **la mémoire** (l'écriture de symboles sur le papier soulage la mémoire, permet de s'arrêter en cours de route et de reprendre plus tard car « le mécanisme conserve en quelque sorte à ma place le souvenir du raisonnement que j'ai fait précédemment[420] »).

La manipulation de symboles extériorise le raisonnement, elle l'objective dans un langage commun qui facilite l'évaluation, la critique, ou l'amélioration. C'est donc à la fois une *prothèse individuelle* et un outil *d'élaboration collective de la connaissance*. Paradoxalement, la « démission de pensée » que certains dénoncent à propos du formalisme, serait au contraire la voie royale pour un *progrès de la pensée.* Pour Whitehead,

> c'est un truisme profondément erroné, répété par tous les manuels et par des gens éminents dans leurs discours, que nous devrions cultiver l'habitude de penser à ce que nous faisons. C'est précisément le contraire qui est vrai. La civilisation progresse en augmentant le nombre d'opérations importantes que nous pouvons effectuer sans y penser[421].

L'idée d'une exonération des humains de tâches pénibles était déjà présente dans l'élaboration de l'algèbre où tout se passe comme si les symboles faisaient l'opération tout seuls. La création ultérieure d'automates ou d'ordinateurs s'inscrit dans la continuité de l'algèbre et remplit la même fonction, c'est-à-dire nous « décharger du labeur intellectuel en confiant aux signes le soin de penser à [notre] place[422]. »

2.3.2. L'algèbre comme *ars inveniendi*

Pour Leibniz et les mathématiciens de son époque, le formalisme n'a pas la réputation désastreuse que lui attribuent ceux qui s'arrêtent sur la soumission absolue à des règles et en font une démission de la pensée, une pratique « déshumanisante » requérant l'ablation

420 Laurence Bouquiaux, *op. cit.*
421 Alfred North Whitehead, *An Introduction to Mathematics,* New York, H. Holt, 1911, p. 59 traduit et cité par Laurence Bouquiaux, *op. cit.*
422 Michel Bourdeau, *op. cit.,* p. 68.

temporaire de l'intuition. Bien au contraire, c'est un *ars inveniendi,* un instrument de découverte. Historiquement,

> l'algèbre a longtemps été une discipline plus pratique que théorique, un art plus qu'une science, puisqu'elle est née non pour démontrer des vérités déjà connues, mais pour résoudre des problèmes, et qu'elle ne poursuivait donc d'autre but que d'étendre notre puissance de calcul[423].

Michel Bourdeau insiste sur cette « fécondité de l'algèbre », loin de l'accusation d'être « stérile et vide », « sous prétexte que la pensée symbolique fait abstraction du contenu des expressions[424] ».

La manipulation de symboles mathématiques a une parenté avec l'exercice de l'imagination. Elle délègue aux règles de syntaxe le pouvoir de transformer une formule, ressemblant à la manière dont l'imagination laisse une image en susciter une autre. La suspension d'une partie du sens permet à l'esprit de prendre des chemins de traverse. Dans les deux cas, il y a une relative autonomie de la forme permise par la suspension temporaire de la fonction de dénotation du signe ou de l'image. Dans les deux cas, l'autonomie de la forme autorisée par un écart à la référence peut être féconde, c'est-à-dire apporter une connaissance pertinente par rapport au réel. En d'autres termes, on peut qualifier la manipulation de symboles mathématiques comme une forme d'*imagination artificielle* qui s'appuie sur la prothèse du langage symbolique.

Le langage symbolique supplée de deux façons à l'imagination : d'une part en permettant de travailler sur des objets qu'il est difficile ou impossible d'imaginer (un polygone à cent côtés, un espace à cent dimensions), et d'autre part en faisant émerger de nouvelles formes pertinentes en s'autorisant un jeu sur les formes qui ressemble au fonctionnement de l'imagination. Il serait comme une canne permettant à la pensée d'aller tâtonner au-delà de son horizon, dans les zones qui échappent à « l'œil de la raison ». Écrivant à son tuteur De Morgan, Ada Lovelace s'émerveille de la manière surprenante dont une expression mathématique peut être transformée, lui rappelant « les lutins et les fées » qui prennent une forme à un moment et une autre tout de suite après. Comme eux, les expressions mathématiques ont une capacité rare à se montrer « trompeuses, ardues et captivantes[425] ». Contrairement aux critiques convenues

423 *Ibid,* p. 86.
424 *Ibid,* p. 95.
425 « And by the bye, I may here remark that the curious transformations many formulae can undergo, the unexpected & to a beginner apparently impossible identity of forms exceedingly dissimilar at first sight, is I think one of the chief difficulties in the early part of mathematical studies. I am often reminded of certain

de la pensée formelle qui l'associent à une censure de la créativité, l'expérience des mathématiciens montre que celle-ci peut être l'occasion de l'émergence de formes « nouvelles », ou en tout cas surprenantes pour celui qui manipule les symboles.

2.3.3. La surprise de Turing contre la nouveauté de Lovelace

Bien consciente de la fécondité du formalisme, Ada Lovelace prend toutefois soin de préciser dans son article sur la machine de Babbage, que celle-ci « ne peut rien créer par elle-même. » Lovelace utilise le mot anglais « originate » que nous devons prendre dans le sens d'« être à l'origine » de quelque chose, produire quelque chose de plus que ce qui est dans les rouages ou le programme. « Elle peut faire tout ce qu'*on sait lui commander* d'effectuer. Elle peut *suivre* l'analyse mais ne peut *anticiper* les relations analytiques et les vérités[426] ». Elle ne fait que « *rendre disponible* ce que nous connaissons déjà. »

Nous avons vu qu'Alan Turing mentionne l'argument de Lovelace dans son article de 1950, en le reformulant de la façon suivante : une machine « ne peut rien faire de réellement nouveau[427] ». Avec facétie, Turing décale le problème vers les humains : quand un travail est présenté comme « original », qui peut prétendre que ce n'est pas simplement « la croissance d'une graine plantée par l'enseignement » ou encore « l'effet de l'utilisation de principes bien connus » ? En d'autres termes, en quoi est-ce que nous, les humains, pourrions prétendre que nous « sommes à l'origine » de quoi que ce soit ?

Puis Turing revient aux machines et reformule à nouveau l'argument de Lovelace : « une meilleure variante de l'objection est qu'une machine ne peut jamais 'nous prendre par surprise' [au sens de 'prendre au dépourvu'][428] ». Il lui oppose alors sa propre expérience : « les machines me prennent par surprise [me prennent au dépourvu] très fréquemment.» Cela

---

sprites & fairies one reads of, who are at one's elbow in one shape now, & the next minute in a form the most dissimilar, and uncommonly deceptive, troublesome & tantalizing are the mathematical sprites & fairies sometimes; like the types I have found for them in the world of Fiction (LB170, [Jan.1841], ff. 91r–91v). » cité par Hollings et al. « The Lovelace-De Morgan mathematical correspondence: A critical re-appraisal », *Historia Mathematica*, 2017.

426 « The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we *know how to order it* to perform. It can *follow* analysis; but it has no power of *anticipating* any analytical relations or truths. Its province is to assist us in making *available* what we are already acquainted with. » Ada Lovelace, « Sketch of the analytical engine invented by Charles Babbage, by LF Menabrea, officer of the military engineers, with notes upon the memoir by the translator », *op. cit.*

427 « A variant of Lady Lovelace's objection states that a machine can 'never do anything really new.' » Alan Turing, « Computing Machinery and Intelligence », *op. cit.*

428 « A better variant of the objection says that a machine can never "take us by surprise." », *Ibid.*

s'explique par le fait qu'il ne « fait pas assez de calculs » pour pouvoir prévoir ce que va faire la machine, ou que, s'il fait les calculs, il les fait « à la hâte, d'une manière bâclée, en prenant des risques » et en se basant sur des suppositions. En conséquence, il est souvent pris au dépourvu par le résultat. Turing en conclut qu'on pourrait le sermonner sur ses mauvaises habitudes en tant que mathématicien, mais pas remettre en cause le fait que les machines le surprennent réellement[429].

Pourtant, comme en témoigne la remarque de Lovelace au sujet de sa surprise face aux « lutins et fées mathématiques », son expérience de la fécondité du formalisme devrait la ranger du côté de Turing. De fait, juste après avoir affirmé que la machine de Babbage ne peut rien créer par elle-même, elle remarque qu'en « combinant les vérités » la machine permettra de les voir « sous de nouvelles lumières » et de les étudier avec plus d'acuité[430]. « Il est évident », ajoute-t-elle, que cette nouvelle manière de noter et d'user des vérités mathématiques permettra d'induire « de nouvelles perspectives[431] ». Si Lovelace récuse que la machine de Babbage puisse créer, elle ne s'opposerait pas à l'idée qu'elle puisse amener des surprises, voire constituer un outil d'investigation, au même titre que l'algèbre. A partir du moment où l'on délégue des opérations à un système formel, qu'il s'agisse d'un ordinateur des années cinquante, d'une machine de Babbage, ou simplement de caractères écrits sur le papier, rien de surprenant à ce que le résultat retourné puisse nous surprendre.

Le désaccord entre Turing et Lovelace tient plutôt à l'interprétation de la surprise qu'à la question de sa possibilité. Quand une machine me surprend, puis-je pour autant affirmer qu'elle est « à l'origine » de quelque chose ? Turing, aussi bien que Lovelace, répondraient probablement tous les deux par la négative, mais pour des raisons différentes. Pour Lovelace, une machine peut me surprendre sans être « à l'origine » de quelque chose, mais cette surprise est à distinguer de celle que provoque la créativité humaine – les humains, eux, peuvent effectivement inventer des formes dont ils sont « à l'origine ». Turing apporte une interprétation

429 « Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks. Perhaps I say tomyself, "I suppose the Voltage here ought to be the same as there : anyway let's assume it is." Naturally I am often wrong, and the result is a surprise for me for by the time the experiment is done these assumptions have been forgotten. These admissions lay me open to lectures on the subject of my vicious ways, but do not throw any doubt on my credibility when I testify to the surprises I experience. » *Ibid.*

430 « For, in so distributing and combining the truths and the formulæ of analysis, that they may become most easily and rapidly amenable to the mechanical combinations of the engine, the relations and the nature of many subjects in that science are necessarily thrown into new lights, and more profoundly investigated. » Ada Lovelace, *op. cit.*

431 « It is however pretty evident, on general principles, that in devising for mathematical truths a new form in which to record and throw themselves out for actual use, views are likely to be induced, which should again react on the more theoretical phase of the subject. » *Ibid.*

inverse : si une machine peut me surprendre sans être « à l'origine » de quoi que ce soit, alors peut-être que les humains, eux aussi, ne sont « à l'origine » de rien. Si les machines et les humains peuvent surprendre Turing, c'est en raison de ses attentes erronées. La surprise vient du défaut de connaissance, et non de l'émergence de quelque chose de nouveau depuis la machine ou l'humain. Elle tient plus à l'interprétation des événements qu'aux événements eux-mêmes.

Turing remarque qu'on pourrait lui opposer que lorsqu'il est surpris, la créativité vient « d'un acte créatif de [sa] part, et ne permet de rien attribuer à la machine[432] », mais l'argument pourrait être retourné à l'envoyeur : quand celui-ci *interprète* la surprise de Turing face à la machine comme un acte créatif *de Turing*, est-ce qu'il ne fait pas à son tour un acte créatif ? « Cela vaut sans doute la peine de remarquer que l'appréciation de quelque chose comme surprenant exige autant 'd'acte mental créatif' que l'événement surprenant prenne son origine [originates] d'un homme, d'un livre, d'une machine, ou quoi que ce soit d'autre[433]. » En d'autres termes, si on souhaite « disqualifier » les machines par rapport aux humains en s'appuyant sur l'argument de la créativité, il faudrait d'abord « qualifier » les humains. Or, nous ne savons pas si nous créons effectivement ou si nous ne faisons que jouer à nous surprendre en déjouant nos propres attentes. Qu'est-ce qui nous permet d'affirmer que, plus que les machines, nous serions « à l'origine » de quoi que ce soit ?

Pour Turing, le point de vue selon lequel « les machines ne peuvent pas causer de surprise » est dû à la supposition erronée qu'« aussitôt qu'un fait est présenté à l'esprit toutes les conséquences de ce fait émergent simultanément[434]. » C'est donc l'inverse qui serait vrai : lorsque nous avons un fait à l'esprit, nous sommes incapables de voir ses conséquences. Turing se contente d'énoncer ce fait sans lui donner d'explications. Pourtant, celui-ci peut correspondre à deux types de justifications opposées : ou bien le réel est contingent et quelle que soit la puissance de notre esprit, il serait impossible de percevoir « toutes les conséquences » d'un fait car un même fait peut donner, de façon imprévisible, différentes conséquences ; ou bien le réel est déterminé, et c'est la faiblesse de notre esprit qui est en cause.

Si le réel est contingent, alors tout peut potentiellement être « à l'origine » de quelque chose, il peut, en droit, y avoir émergence de n'importe quoi n'importe quand : les machines

---

432 « I do not expect this reply to silence my critic. He will probably say that surprises are due to some creative mental act on my part, and reflect no credit on the machine », Alan Turing, *op. cit.*

433 « It is perhaps worth remarking that the appreciation of something as surprising requires as much of a "creative mental act" whether the surprising event originates from a man, a book, a machine or anything else. » *Ibid.*

434 « The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. » *Ibid.*

peuvent être « créatives » au même titre que les humains, mais aussi que les casseroles et les cailloux. Il ne s'agit donc pas de rendre les machines créatives, puisque tout est déjà créatif. Par contre, si le réel est déterminé, il n'y a jamais de création, uniquement des transformations selon des règles et nous ne faisons que jouer à nous surprendre grâce à notre ignorance. Dans un cas comme dans l'autre, il n'y a pas de différence entre humain et machine. La bonne question n'est donc ni de savoir si les machines peuvent être à l'origine de quelque chose (argument de Lovelace), ni de s'accorder sur le fait qu'elles peuvent nous surprendre (réponse de Turing) mais plutôt : « est-ce que la notion de nouveauté a un sens ? ». Plus précisément, est-ce qu'il est possible de concilier une telle notion avec un esprit scientifique qui souscrit au principe de raison selon lequel tout a une cause ?

Avant de revenir sur ces problèmes, nous retenons ce parti pris par Turing : les « inventions » de l'esprit humain (ou ce qui apparaît comme des inventions à cause de la suprise qu'elles provoquent), ne constituent pas une objection valable au projet d'imitation de la pensée par un système formel matérialisé. Cela présuppose un glissement et une extension du domaine de la pertinence de la pensée formelle : celle-ci n'est plus seulement considérée comme *un outil de travail* sur lequel la pensée s'appuye et qui implique de s'aveugler momentanément en échange d'un résultat fiable. Elle devient un *moyen de décrire* ce que la pensée effectue (glissement) aussi bien quand elle calcule que quand elle ne semble pas être en train de calculer (extension).

En prenant cet outil où la pensée s'aveugle *momentanément* comme un média capable de décrire *tous les moments de la pensée*, n'y a-t-il pas un risque d'aboutir à une image tronquée de la pensée ? Autrement dit, en utilisant les moyens de la « pensée aveugle » pour décrire *toute* la pensée, comment ne pas aboutir à une « image de la pensée » où celle-ci n'est pas momentanément, mais *toujours* aveugle ?

2.4. Attaques contre l'intuition

2.4.1. « Tous les pas sont glissants » : peut-on formaliser l'ensemble de la pensée ?

Dans la lignée du postulat de Hobbes (« la raison n'est rien d'autre que le calcul[435] ») et du rêve de Lulle – inventer un « art » avec pour finalité de « répondre à toutes les questions, pourvu qu'on sache bien la signification de chacun de ses termes[436] » –, Leibniz souscrit à l'idée d'une extension du formalisme à l'ensemble de la pensée. D'une part, il serait dommage de réserver les bénéfices du formalisme à l'algèbre et à la géométrie, et d'autre part, nous utilisons déjà un système formel – moins bien conçu – pour les pensées ordinaires. Dès que nous manipulons des notions complexes, nous le faisons « par l'intermédiaire de signes qui sont, pour notre entendement fini, plus faciles à manipuler que les idées elles-mêmes ». Si on l'associe généralement à l'algèbre et aux mathématiques, elle est « en réalité […] en œuvre dès que nous pensons une notion complexe – c'est-à-dire dès que nous pensons, toute pensée discursive (*ratiocinatio*) consistant en une utilisation de signes[437]. » Donc, « la plus grande part de la connaissance humaine » repose sur l'usage de symboles, et nous ne pourrions pas penser comme nous le faisons sans les symboles. Autrement dit, « la plus grande part » de notre pensée s'effectue à travers la médiation de signes[438]. Cependant, seules les mathématiques ont développé un formalisme assez précis pour qu'on puisse passer par une forme de « pensée aveugle ». Dans les autres domaines, « tous les pas sont glissans » et il est présomptueux de « donner des démonstrations en matière de physique, de métaphysique, de morale, et même en politique, en jurisprudence et en médecine[439] ». Pour Leibniz, si nous disposions d'un système

435 Thomas Hobbes, *Léviathan,* Paris, Garnier Flammarion, 2017, chapitre cinq.

436 Josep Rubio, *Raymond Lulle, le langage et la raison : une introduction à la genèse de l'Ars,* Paris, Vrin, 2007. Le premier ouvrage de Leibniz est un commentaire de l'œuvre de Lulle.

437 Mildred Galland-Szymkowiak, *op. cit.*

438 *Ibid.*

439 « [Q]uoyque beaucoup de tres habiles gens, surtout de nostre siecle, ayent pretendu de nous donner des demonstrations en matiere de physique, de metaphysique, de morale, et même en politique, en jurisprudence et en medecine : neantmoins ou ils se sont trompés, à cause que tous les pas sont glissans, et qu'il est difficile de ne pas tomber, lorsqu'on n'est pas guidé par quelques directions ou quand même ils ont rencontré, ils n'ont pas pu faire recevoir leur raisonnemens de tout le monde ; par ce qu'il n'y a pas encor eu moyen d'examiner

de signes suffisamment fiable et précis, nous pourrions appliquer le même formalisme qu'en mathématiques aux autres domaines du savoir.

> si l'on pouvoit trouver des caractères ou signes propres à exprimer toutes nos pensées, aussi nettement et exactement que l'arithmetique exprime les nombres, ou que [l' algèbre] l'analyse géométrique exprime les lignes, on pourroit faire en toutes les matières autant qu'elles sont sujettes au raisonnement tout ce qu'on peut faire en Arithmétique et en Geometrie[440].

Leibniz rêve d'un système de signes non équivoques, une langue universelle (« caractéristique universelle »), avec des règles de syntaxes telles que le fait de les respecter suffirait à se prémunir de l'expression d'idées fausses.

> Ceux qui écriront en cette langue, ne se tromperont pas pourveu qu'ils evitent les erreurs de calcul, et barbarismes, solecismes et autres fautes de grammaire et de construction. De plus cette langue aura une propriété merveilleuse, qui est de fermer la bouche aux ignorans. Car on ne pourra pas parler ny ecrire en cette langue que de ce qu'on entend : ou si on ose le faire, il arrivera de deux choses une, ou que la vanité de ce qu'on avance soit manifeste à tout le monde, ou qu'on apprenne en écrivant ou en parlant[441].

Du seul fait de son système de règles, une telle langue serait débarrassée d'énoncés mal formulés. Elle simplifierait la philosophie en évitant tout débat là où il n'y a qu'une divergence de définitions. Si une telle langue était possible, cela soulagerait la réflexion comme la manipulation symbolique soulage la recherche mathématique.

> Car toutes les recherches qui dependent du raisonnement se feroient par transposition de ces caracteres, et par une espèce de calcul ; ce qui rendroit l'invention des belles choses tout a fait aisée. Car il ne faudroit pas se rompre la teste autant qu'on est obligé de faire aujourd'huy […][442].

Enfin, avantage non négligeable, les débats de fonds seraient remplacés par des considérations sur la forme. En cas de désaccord, il suffirait de contrôler la pertinence formelle des énoncés

---

les raisonnemens par quelques preuves aisées dont tout le monde fut capable. » Gottfried Wilhelm Leibniz, « Préface à la science générale », in Louis Couturat (ed.), *Opuscules et fragments inédits de Leibniz : extraits des manuscrits de la Bibliothèque royale de Hanovre*, Paris, Félix Alcan, 1903, p. 155.

440 *Ibid.*

441 *Ibid*, p. 156.

442 *Ibid*, p. 155.

pour trouver une entente. « Et si quelqu'un doutoit de ce que j'aurois avancé, je luy dirois : contons, Monsieur, et ainsi prenant la plume et de l'encre, nous sortirions bientost d'affaire[443]. » L'histoire de la pensée a retenu ce rêve de Leibniz sous l'interjection « calculemus ! » (« calculons! »), c'est-à-dire : présupposons que si les choses sont bien formulées, nous ne pouvons qu'être d'accords, et reformulons correctement nos énoncés de façon à déterminer lequel d'entre nous s'est égaré.

Le projet de caractéristique universelle est celui d'une langue si rigoureusement formée qu'elle serait capable de *mettre tout le monde d'accord,* autrement dit d'émanciper la recherche de la vérité de la variabilité individuelle. Leibniz reproche à Descartes de fonder sa méthode sur une évidence de l'intuition (ce qui est « clair et distinct ») qui n'est pas toujours partagée. Ce qui est évident pour l'un ne l'est pas forcément pour l'autre. Tandis qu'en s'appuyant sur des formes communes rigoureusement définies, le travail de la pensée peut devenir un projet collectif[444].

Avec le projet de caractéristique universelle, Leibniz ne décrit pas comment nous pensons, mais comment nous *devrions* penser. Mais le projet implique de souscrire à une certaine image de la pensée. Se laisser aller à rêver avec Leibniz nous amène à partager les mêmes présupposés optimistes :

(1) il y a une vérité universelle ;

(2) chacun peut la « voir » grâce à la raison ;

(3) les désaccords proviennent d'un mauvais usage de la raison ;

(4) celui-ci peut être corrigé en s'appuyant sur la pensée formelle.

Autrement dit, le formalisme agit comme une paire de lunettes pour « l'œil de la raison » qu'est l'intuition et les philosophes seraient – comme Spinoza en son temps – avant tout des polisseurs de lentilles.

Mais l'optimisme de Leibniz ne fait pas l'unanimité. Chacun des présupposés avancés peut être contesté : qu'est-ce qui garantit que ce que nous « voyons » est commun et que nous contemplons un même paysage – qu'il y a bien une vérité universelle ? Est-ce que ce paysage

443 *Ibid*, p. 156.

444 C'est le sens du commentaire qu'en donne Laurence Bouquiaux : « Le formalisme permet de transformer une opération subjective et dont l'appréciation ne peut être faite qu'en première personne – c'est moi et moi seul qui sais si je suis attentif lorsque, par exemple, j'effectue un calcul mental – en un exercice collectif – nous pouvons, ensemble, vérifier qu'il n'y a pas d'erreurs d'inattention dans le calcul écrit. C'est bien là le sens du célèbre « *calculemus !* ». Si vous estimez que je me suis trompé, prenons nos plumes et parcourons ensemble les différentes étapes du raisonnement. C'est ainsi que Leibniz a pu voir dans sa caractéristique un « juge des controverses » qui peut mettre fin aux disputes, et cela bien mieux que ne peut le faire le critère cartésien du clair et du distinct – car ce qui apparaît évident à l'un n'apparaît pas toujours évident à l'autre. Les discussions qui portent sur les signes ont bien plus de chances d'aboutir que celles qui portent sur les idées. » Laurence Bouquiaux, *op. cit.*

préexiste et nous le découvrons, ou bien est-ce que nous l'inventons ? De quelle nature est cette « perception » par la raison, et jusqu'à quel point la comparaison avec la vision est-elle pertinente ? Qu'est-ce qui garantit qu'une perception de la raison est juste ? Qu'est-ce qui fonde notre confiance dans le formalisme ? Celui-ci est-il une *correction* de notre vision de façon à mieux « percevoir » des objets préexistants ou bien une *construction* d'objets communs ? S'il est efficace en mathématique, cela implique-t-il qu'il le soit pour traiter des concepts philosophiques ?

2.4.2. Le programme de Hilbert : l'intuition « à la niche ! »

Les mathématiciens de la fin du dix-neuvième siècle se sont inscrits dans la lignée du formalisme de Leibniz en s'efforçant de réformer les mathématiques pour « ne rien admettre […] qui n'ait été prouvé par une démonstration valable ». David Rabouin cite « l'anathème » jeté par Hilbert et Pasch « contre leur grand ancêtre, Euclide », coupable de fonder ses propositions sur les sens et non sur des règles de déduction. Pour Pasch, la géométrie doit s'émanciper de la signification et ne reposer que sur des procédés de raisonnement indépendants :

> Si la géométrie veut devenir une science véritablement déductive, il faut que ses procédés de raisonnement soient indépendants de la signification des concepts géométriques, comme ils sont indépendants des figures ; seules les relations imposées à ces concepts par les postulats et les définitions doivent intervenir dans la déduction[445] .

Et David Rabouin de commenter « L'intuition à la niche[446] ! » Pour autant, cela ne signifie pas qu'elle soit éliminée. Elle conserve une « niche », celle des « propositions premières : postulats et définitions » ou, pour reprendre les termes de Leibniz, des « vérités premières ou bien des hypothèses ». Pour celles-ci, il a bien fallu admettre, faute de mieux, le recours à l'intuition. Aussi, le projet formaliste n'a pas tant consisté à éradiquer l'intuition qu'à la déplacer des mathématiques vers les métamathématiques.

445 Moritz Pasch, cité par David Rabouin, « Penser comme un pied », *op. cit.*
446 David Rabouin, *Ibid.*

> [Hilbert] ne se proposait pas tant d'éliminer l'intuition que d'en déplacer le champ d'application : vers la perception sensible, bien sûr, en quoi il préparait la mécanisation du calcul, mais aussi vers un nouveau domaine, la métamathématique, où le raisonnement s'effectuait de manière informelle, comme par le passé[447].

L'approche purement formelle qui « refus[e] de prendre en compte le sens des symboles » est méthodologique. Cela est « destiné à assurer aux mathématiques usuelles un fondement inébranlable[448]. » Mais il ne s'agit pas pour autant d'opérer une démission de la pensée. L'exercice de l'intuition est simplement déplacé vers le travail sur les fondements : « Si la mathématique a été réduite à un ensemble de formules, ce n'est que pour faciliter la tâche du métamathématicien et, du point de vue plus élevé qui est le sien, la pensée intuitive retrouve tous ses droits[449]. »

2.4.3. La machine de Turing : matérialiser et unifier les prothèses de la pensée

L'histoire du calcul n'a pas consisté à déléguer progressivement le calcul mental vers un support matériel, mais plutôt à élaborer et raffiner des supports matériels qui ont toujours accompagné l'exercice du calcul : compter sur ses doigts, recourir à des petits cailloux (« calculis ») ou encore aux abaques et bouliers, auxquels ont succédé les calculateurs puis les ordinateurs[450]. Bien avant l'invention des calculateurs, l'exercice de la pensée formelle à l'aide de l'écriture est déjà une *matérialisation* du calcul puisqu'une partie de l'effort est transférée de l'intellect vers le sensible – les yeux et la main[451].

S'ils sont indissociables d'opérateurs humains, les supports matériels du calcul requièrent de moins en moins d'attention et de réflexion de leur part. Grâce à l'invention de l'algorithme pendant l'Antiquité, la pensée entame un retrait du calcul qui s'accentue avec

447 Michel Bourdeau, *op. cit.*, p. 122.
448 *Ibid.*
449 *Ibid*, p. 26.
450 « de nombreux indices tirés des œuvres littéraires et des fouilles archéologiques nous révèlent que la pratique du calcul était outillée par des instruments matériels variés permettant d'aller au-delà des possibilités offertes directement par notre corps, à savoir le calcul mental et le calcul avec les doigts. » Dominique Tournès, « Perspectives historiques sur les abaques et bouliers », *MathémaTICE*, 2016, 51. hal-01480067.
451 Le formalisme « permet de résoudre les problèmes à moindre frais, sans exiger de maintenir constamment un haut degré d'attention, parce qu'il autorise le recours au sensible. Pour reprendre les termes de Whitehead, une partie du travail qui devrait autrement être faite par l'intellect est réalisé par les yeux. » Laurence Bouquiaux, *op. cit.*

l'amélioration des supports. Au dix-huitième siècle, l'application de la division du travail est facilitée par le fractionnement en tâches élémentaires qu'implique la notion d'algorithme – le travail est déjà divisé, il faut simplement le répartir – et marque un tournant en séparant nettement les moments où il faut encore réfléchir (mise en équation et interprétation des résultats) des moments où il faut opérer « aveuglément ». Dans l'usine à calculer de Gaspard de Prony[452], seuls les mathématiciens supervisant les opérations ont conscience du sens de celles-ci. Tous les autres employés se contentent de faire des additions et des soustractions, ou de comparer les résultats pour vérifier qu'ils sont égaux.

Lorsque Charles Babbage découvre une description de l'usine il est immédiatement frappé par le fait que les opérations des employés sont si simples qu'elles peuvent être mécanisées (les calculateurs élémentaires existent depuis le dix-septième siècle grâce aux travaux de Leibniz et Pascal). Au cours de la fabrication de ce qu'il baptise le « Difference Engine », il a une nouvelle idée qui peut s'interpréter rétrospectivement comme le « pas suivant » en direction de l'ordinateur. Ayant eu vent de l'utilisation de cartes perforées pour réagencer les métiers à tisser, Babbage réalise qu'au lieu de construire une machine correspondant à une seule suite d'instructions (produire une table de logarithme), il pourrait construire une machine qui peut se modifier pour effectuer différentes suites d'instructions – de la même façon que les métiers à tisser sont reprogrammés pour changer de motif. Il ne s'agit plus seulement de mécaniser les opérations des employés de l'usine à calculer, mais également une partie de celle des mathématiciens : les tâches consistant à donner des instructions et à passer d'une suite d'instructions à une autre.

Ainsi, les supports de calcul ont évolué selon plusieurs tendances qui se sont renforcées mutuellement :

(1) déléguer les opérations de l'intellect vers le sensible assisté par un support matériel ;

(2) simplifier les opérations à effectuer ;

(3) affranchir le support matériel des interventions humaines ;

(4) généraliser, voire universaliser les opérations pour que le moins de machines possibles puissent effectuer le plus de calculs.

452 Pendant la Révolution, Gaspard de Prony est chargé de produire des tables de logarithmes utiles aux marins, commerçants et artilleurs. Il recrute plusieurs centaines de chômeurs (dont des perruquiers) ne connaissant que le calcul élémentaire ainsi que quelques mathématiciens et organise leur travail suivant les principes décrits par Adam Smith dans le fameux passage de la manufacture d'épingle. Margaret Bradley, « Gaspard-Clair-François-Marie Riche de Prony (1755-1839), Constructeur de ponts », in *Bulletin de la SABIX*, « Regards sur des carrières de polytechniciens au XIXe siècle », 48 | 2011.

A la fin du dix-neuvième siècle et au début du vingtième, de nombreux travaux prolongent ces évolutions, comme ceux de Boole (décomposer le calcul en un minimum d'opérations élémentaires, nécessitant un minimum de signes) ou de Shannon (matérialiser les opérations élémentaires). C'est dans l'article de Turing de 1936[453] qu'elles trouvent leur point culminant.

Les travaux de Gödel ont établi que, dans un système formel élémentaire (arithmétique de Peano ou théorie des ensembles), il existe des propositions indécidables, c'est-à-dire des propositions pour lesquelles il n'est pas possible de prouver qu'elles sont consistantes avec les axiomes du système. Alan Turing se propose de prolonger ce résultat en montrant qu'il n'est pas non plus possible d'établir une procédure qui permette de trancher a priori sur le statut d'une proposition (est-elle décidable ou non ?). Pour faire sa démonstration, il a besoin de formaliser cette notion de procédure, ou d'algorithme, de la manière la plus élémentaire. Il la réduit à « trois opérations fondamentales : lire et écrire un symbole, parcourir la feuille sur laquelle ceux-ci sont inscrits, passer à une autre opération[454]. » En décomposant la notion d'algorithme en des éléments aussi simples, Turing isole « ce qu'il y a tout lieu de considérer comme les atomes, les éléments irréductibles du calcul[455] ». Il atteint un point de simplicité extrême : deux signes (0 et 1) et seulement trois opérations ; mais aussi d'universalité : un seul modèle de machine peut exécuter n'importe quel algorithme. Pour pouvoir établir que « tout, même dans le champ de la mathématique, n'est pas calculable[456] », Turing a été amené à décrire « l'existence d'une machine qui calcule tout ce qui est calculable[457] ». L'idée de machine universelle, qui n'est qu'un adjuvant à sa démonstration, aura une postérité considérable, inspirant aussi bien l'invention des réseaux de neurones que celle de l'ordinateur.

Alors que l'article établit des limites au formalisme (il n'existe pas d'algorithme qui permette de savoir s'il existe un algorithme pour arriver à un résultat donné), un de ses éléments, la machine de Turing, réactualise le rêve de Leibniz d'étendre le formalisme aux affaires humaines. Tout comme la « caractéristique » de Leibniz, la machine de Turing se base sur un atomisme, une réduction de tous les termes à des primitives binaires assemblées selon des règles simples de composition. À la différence de la caractéristique, elle ne formalise pas les concepts mais les procédures. Par contre, elle est effectivement universelle : tous les algorithmes sans exception peuvent être formalisés de cette façon. Avec l'invention de l'ordinateur, ce qui est

453 Alan Turing, « On Computable Numbers, With an Application to the Entscheidungsproblem », Proceedings of the London Mathematical Society, Ser. 2, vol. 42, 1937.
454 Michel Bourdeau, *op. cit.*, p. 43.
455 *Ibid.*
456 *Ibid*, p. 55.
457 *Ibid.*

une universalité théorique ouvre la possibilité d'une universalité pratique : si la machine de Turing peut effectuer tous les algorithmes, peut-être l'ordinateur pourra-t-il effectuer toutes les procédures humaines ? Voilà qu'un système formel épuré, par le biais de sa matérialisation en ordinateur, s'immisce dans les affaires humaines, ravivant le rêve de Leibniz d'étendre le formalisme des mathématiques au reste de la pensée.

L'histoire de l'évolution des supports de calcul se confond rétrospectivement avec la matérialisation progressive *d'une prothèse unique*, l'ordinateur et son modèle idéal qu'est la machine de Turing. Avec cette dernière, les éléments du calcul sont devenus si simples qu'un agencement matériel peut les effectuer. Le calculateur n'a plus besoin de *faire la machine,* il préférera désormais la *fabriquer*. Comme souvent, la lecture retrospective des éléments donne une impression de téléologie. Tout se passe comme si s'étaient progressivement imbriquées les pièces d'un puzzle correspondant à l'invention de l'ordinateur, et les concepts sont si bien ancrés qu'on ne voit pas comment on pourrait en avoir une théorie différente. Cela donne l'impression que *l'idée de l'ordinateur* était déjà présente bien avant (certains auteurs remontent jusqu'à Lulle) et qu'elle n'a eu qu'à s'incarner au fur et à mesure des travaux de différentes lignées de chercheurs. C'est un cas, pour reprendre l'expression de Bergson, « d'illusion retrospective du vrai » négligeant les multiples contingences qui jalonnent l'histoire de son invention. Sans qu'il y ait téléologie, l'activité artificielle (manipulation de symboles mathématiques selon des règles formelles) à laquelle les humains s'adonnaient devient une véritable prothèse extérieure et autonome – gagnant au passage une efficacité sans précédent. C'est un tour nouveau – ou une conséquence surprenante – que la machine de Turing, puis l'ordinateur, donnent au rêve de Leibniz. Pour le philosophe de Göttingen, il s'agissait d'appliquer le formalisme des mathématiques à la pensée non mathématique. En 1950, pour Alan Turing il s'agit plutôt de se demander si, à force de raffinement, la prothèse pourrait se substituer à l'organe qu'elle assiste : un ordinateur pourrait-il penser ?

2.4.4. GOFAI : éliminer l'intuition

Avec le projet formaliste, la notion d'intuition a été placée sur le banc des accusés. On lui a reproché son inconstance, sa variabilité, les difficultés à la décrire et à justifier ses propositions. Mais elle a gardé un droit de cité grâce aux métamathématiques. Descendant indirect du programme de Hilbert, via les travaux de Turing, le projet d'intelligence artificielle n'en retient

que le premier moment : exclure tout recours à l'intuition lors de la manipulation de symboles. La volonté de rigueur qui aboutit chez les mathématiciens a une séparation entre pensée formelle et pensée intuitive devient chez les informaticiens une volonté d'exclusion radicale de l'intuition. Avec l'hypothèse de Dartmouth et le projet d'intelligence artificielle symbolique, a lieu une condamnation définitive : autant l'éliminer. Quand le projet d'intelligence artificielle fait, avec la conjecture de Dartmouth, l'hypothèse que toute la pensée peut se décrire sous la forme d'un algorithme, il choisit de ne considérer *que les propositions calculables*, ce qui est une autre façon de « ne rien admettre […] qui n'ait été prouvé par une démonstration valable ». Le geste fondateur de l'IA revient à « identifier la pensée symbolique à la pensée », ce qui « équivaut alors à nier l'existence d'une quelconque intuition[458]. » En d'autres termes, le projet d'intelligence artificielle présuppose que l'intuition n'est qu'un leurre. Elle ne correspond qu'aux moments de pensée pour lesquels les calculs n'ont pas encore été explicités. Les moments que nous qualifions d'intuition devront être décrits sous la forme de manipulations de symboles et révéler *in fine* leur nature algorithmique. Ce que nous appelons l'intuition ne serait que *la partie du fonctionnement de la pensée que l'on ne comprend pas encore*. Chaque progrès de l'intelligence artificielle, révélant le fonctionnement symbolique de facultés autrefois considérées comme intuitives, sera interprété comme un nouveau pas vers la destitution de l'intuition. Souscrire à la conjecture de Dartmouth et à la possibilité de la fabrication d'une intelligence artificielle générale revient à postuler qu'on peut se passer complètement du concept d'intuition et le remiser au hangar des superstitions d'antan.

> C'est pourquoi cette notion, étrangère à l'IA, est aussi celle qui permet de donner, de sa thèse centrale, la version la plus à même d'indiquer ce qui, philosophiquement, y est en cause : il s'agit d'établir que, connaissance confuse de ce qui peut être pensé distinctement, l'intuition est une hypothèse inutile[459].

L'évolution des prothèses de la pensée vers le modèle universel de l'ordinateur peut être interprétée comme un argument en faveur d'un universalisme de la pensée elle-même. Les ordinateurs ayant tous les mêmes règles de fonctionnement, il est aisé de s'imaginer que nos pensées individuelles reposent sur des principes communs. Tout comme une machine de Turing peut effectuer *n'importe quel* algorithme, peut-être que ces principes s'appliquent à l'ensemble de la pensée. « Grisés par ces résultats inattendus [du formalisme], certains en sont venus à

458 Michel Bourdeau, *op. cit*, p. 12.
459 *Ibid.*

concevoir la pensée symbolique comme susceptible d'une extension indéfinie, qui à la longue rendrait superflu tout autre mode de connaissance[460]. » Pour l'intelligence artificielle symbolique (ou GOFAI pour *good old fashioned AI*), le rêve de Leibniz n'a pas été un souhait à réaliser mais le présupposé sur lequel se fonde la recherche. Le système formel par lequel s'effectue la pensée est moins à inventer qu'à découvrir. La « pensée aveugle » n'est pas un moment isolé du raisonnement que l'on se force à suivre pour s'assurer une rigueur adéquate et provoquer de nouvelles découvertes, c'est *toute la pensée* qui est, en deça des apparences, déjà une manipulation de symboles.

L'école symbolique ne se prononce pas sur le statut à attribuer aux moments de pensée qui semblent échapper à la formalisation. Jusqu'au renouveau connexionniste, les chercheurs en intelligence artificielle n'évoquent pas l'intuition. Bourdeau remarque « le silence total observé à son propos par l'IA et les écoles philosophiques dont elle est indirectement issue[461] ». Pourtant, « bien que la nature de l'intuition pose des problèmes redoutables, on ne voit pas au nom de quoi il serait interdit d'en parler[462]. » De plus, pour rester fidèle à la promesse d'exhaustivité formulée à Dartmouth – décrire *toutes* les facettes de l'intelligence – il faudrait faire un sort à l'intuition.

Le parti pris de l'école symbolique est de décrire toute la pensée comme si l'intuition n'existait pas, en comptant sur le succès de l'entreprise pour valider après coup le présupposé implicite que le concept d'intuition n'est pas pertinent. Mais après avoir escamoté la notion d'intuition pour lui préférer le formalisme, l'école symbolique s'est trouvée incapable de prendre en compte certaines caractéristiques fondamentales de la pensée. En « identifi[ant] pensée et pensée symbolique », la GOFAI se retrouve à faire « purement et simplement l'économie du sens[463]. » Avant d'examiner la proposition de l'école connexionniste, il nous faut mettre en évidence dans quelle mesure le concept résiste aux tentatives d'élimination – montrer pourquoi, malgré tous ses défauts, il incombe de prendre la notion d'intuition au sérieux.

460 *Ibid*, p. 82.
461 *Ibid.*
462 *Ibid.*
463 « Il s'agit de savoir si ce qui précède autorise ou non à franchir un dernier pas et à décider de faire purement et simplement l'économie du sens. C'est, comme chacun sait, le point de vue de l'IA, qui identifie pensée et pensée symbolique, et ne veut voir dans le sens qu'un sous-produit de la syntaxe. **»** Michel Bourdeau, *op. cit.*, p. 102.

2.5. Peut-on se passer de l'intuition ?

2.5.1. Un fait : l'expérience mathématique

Un projet qui viserait à éliminer l'intuition dispose de raisons légitimes : l'intuition varie d'une personne à l'autre, elle ne dispose pas de moyens pour fonder ou justifier ce qu'elle propose, elle est trop fragile pour que l'édifice de la connaissance commune puisse reposer sur elle. Pour autant, peut-on effectivement s'en passer ? Michel Bourdeau répond par la négative. Il mentionne les récits de mathématiciens et de logiciens pour qui elle occupe une place cruciale, notamment chez Gödel[464]. Pour Pierre Cassou-Noguès, « c'est une hypothèse sur laquelle Gödel revient fréquemment dans ses cahiers : l'existence d'un organe de la raison, d'un organe de la perception qui n'est pas tourné vers le domaine sensible[465] » et qui est comme un « œil mathématique ». Gödel mentionne « un fait psychologique de l'existence d'une intuition » qui n'est pas moins factuel que la perception sensible :

> Nous avons quelque chose comme une perception des objets de la théorie des ensembles. Je ne vois pas de raison pour avoir moins de confiance dans cette espèce de perception, c'est-à-dire dans l'intuition mathématique, que dans la perception sensible[466].

À la suite de Gödel, « nombre d'auteurs, se réclamant souvent de Kant ou de Husserl, n'hésitent plus aujourd'hui à reconnaître l'existence d'une intuition mathématique[467] ».

464 « Un platoniste comme lui ne pourrait pas affirmer l'existence d'un ciel des idées, où trônent les objets mathématiques, s'il n'y avait accès par une faculté qui, sans être nécessairement saisie immédiate, est du moins l'analogue, pour l'abstrait, de ce qu'est la perception sensible pour le monde concret dans lequel nous vivons. » *Ibid*, p. 105.

465 Pierre Cassou-Noguès, *Les démons de Gödel, logique et folie*, Paris, Seuil, 2012, p. 86.

466 Kurt Gödel, in Feferman, Dawson et al. (eds.), *Kurt Gödel, Collected Works, Volume II, Publications 1938-1974*, Oxford, Clarendon Press, 1990, p. 268.

467 Michel Bourdeau mentionne les deux premiers chapitres de J. Largeault, *Intuition et intuitionnisme*, Vrin, 1993. ainsi que Ch. Parsons, « Mathematical Intuition », *Proceedings of the Aristotelian Society*, 80 (1979-1980), 145-168 ; R. Tieszen, *Mathematical Intuition*, Kluwer, 1989. ou encore M. Detlefsen (éd.), *Proof and Knowledge in Mathematics*, Routledge, 1992, p. 208-233.

2.5.2. Historiquement, l'intuition précède le formalisme

David Rabouin remarque qu'avant les efforts de formalisation des dix-septième et dix-neuvième siècles, l'absence de règles précises n'a pas empêché les mathématiciens de produire des résultats pertinents.

> Tout irait bien si les mathématiques anciennes étaient fausses. Mais voilà, aucun mathématicien n'irait se risquer à dire que les théorèmes d'Euclide sont faux ! Mal démontrés, peut-être. Imprécis, voire peu intéressants, éventuellement. Mais faux, non, jamais ! D'ailleurs, à tout prendre, il y a moins de faussetés dans les mathématiques d'Euclide, Archimède et Apollonius réunis que dans Bourbaki. Bref, la mathématique n'a jamais eu besoin d'être *wirklich deduktiv* pour être vraie[468].

Le formalisme ne serait donc pas indispensable, puisque les mathématiciens ont pu s'en passer pendant une si longue période. À cela s'ajoute qu'on ne trouve pas plus d'erreurs dans les mathématiques grecques qui reposent sur l'intuition, que dans des travaux plus récents qui s'appuyent sur le formalisme. Cependant, avec l'élaboration du langage et des règles mathématiques, la rigueur a fini par se confondre avec l'emploi du formalisme. Si bien qu'aujourd'hui nous ne comprenons plus comment les mathématiciens antiques et médiévaux ont pu faire pour ne pas se tromper, alors qu'ils ne disposaient pas de l'édifice formel élaboré ultérieurement.

> quelle que soit la critique qu'on fera porter sur le caractère 'intuitif' des savoirs antérieurs, on ne fera que renforcer le mystère : comment se fait-il que cela ait fonctionné ? Comme se fait-il que l'intuition, qu'on aimerait chasser ou du moins cantonner à un rôle minimal au motif qu'elle est trompeuse, ne nous ait justement *pas* trompé[469] ?

Le fait est pourtant là : la confiance en l'intuition et l'absence de règles précises a pu produire de bons résultats. Un dispositif aussi ancien que la géométrie euclidienne n'a pas produit de résultat faux et est resté stable pendant près de 2000 ans[470].

---

468 David Rabouin, « Penser comme un pied », *op. cit.*

469 *Ibid.*

470 David Rabouin reprend une remarque de Kenneth Manders, « The Euclidean Diagram » in P. Mancosu, ed., *The Philosophy of Mathematical Practice,* Oxford, Oxford University Press, 2008, p 80-133 : « accordons, dit Manders, que la mathématique ne soit pas 'rigoureuse' avant une certaine époque, encore nous faudrait-il expliquer que des dispositifs anciens, comme la géométrie euclidienne, aient été stables (pendant près de 2000

2.5.3. Les limites du formalisme au secours de l'intuition ?

Un récit ne vaut pas une preuve et l'expérience des mathématiciens ne suffisent pas pour asseoir la prévalence de l'intuition. Diverses tentatives ont été faites pour combler ce vide en utilisant les limites du formalisme comme argument en faveur de la nécessité de l'intuition. En particulier, les théorèmes de Gödel apportent « une réfutation [du projet formaliste] qui établit que toute tentative pour éliminer l'intuition est vouée à l'échec. » En effet, « la suite de théorèmes par lesquels Gödel mettait brutalement un terme aux espoirs de Hilbert » a permis de « montrer les limites du formalisme » puisque « dans l'arithmétique formelle, à l'aide de laquelle ce dernier comptait mettre les mathématiques à l'abri des contradictions, il existe des propositions indécidables, et le domaine du vrai ne se confond donc pas avec celui du démontrable », ce qui laisse penser qu'il est possible d'en déduire « la nécessité de l'intuition[471] ».

Les théorèmes d'incomplétude permettent-ils de réfuter l'ambition du projet d'intelligence artificielle ? C'est l'interprétation que Gödel a pu avoir de ses propres résultats[472], y voyant même un argument en faveur de l'irréductibilité de la pensée à la matière[473]. Gödel ne critique pas directement la conjecture formulée à Dartmouth mais discute la thèse de Turing au sujet de l'existence de procédure finies non mécaniques[474]. Pour lui, de telles procédures

ans pour la mathématique d'Euclide!) et qu'ils n'aient pas, ou du moins pas plus que d'autres, produit de résultat faux. »

471 Michel Bourdeau, *op. cit.,* p. 82.

472 « Gödel, pour sa part, y voyait une réfutation de ces explications mécaniques de l'esprit, dont l'IA n'est jamais qu'une variante. Ajoutons que le réalisme du logicien autrichien l'a conduit à reconnaître la nécessité de l'intuition : si les objets mathématiques ne sont pas une pure construction de l'esprit mais existent indépendamment de nous, il faut, pour les connaître, une faculté analogue à cette intuition sensible par laquelle les objets physiques nous sont donnés. Préfère-t-on parler en termes de jugements qu'on aboutirait à la même conclusion : le domaine du connaissable excède celui du démontrable, et seule l'intuition peut nous donner accès aux vérités qui ne sont pas des théorèmes. » *Ibid,* p. 82. Voir également Pierre Cassou-Noguès, *Les démons de Gödel, op. cit.*

473 « L'esprit étant capable de concevoir des vérités qui échappent aux machines, il faudrait donner raison aux adversaires du matérialisme et admettre que la pensée n'est pas réductible à l'étendue. Gödel, le premier, a tiré ce genre de conséquences. Le mécanisme en biologie n'était pour lui qu'un préjugé de notre époque. La probabilité pour passer, par les seules lois de la physique, d'une distribution aléatoire de particules élémentaires à un être organisé comme le corps humain était beaucoup trop faible pour mériter d'être prise en considération, et il croyait même possible de démontrer empiriquement que l'esprit n'est pas réductible à la matière. » Michel Bourdeau, *op. cit.,* p. 109.

474 Pierre Cassou-Noguès, « Gödel et la thèse de Turing », *Revue d'histoire des mathématiques,* n°14, 2008, p. 77-111, et Inês Hipolito, « Gödel on the mathematician's mind and Turing Machines », *E-Logos Electronic Journal for Philosophy*, n°22, 2014.

pourraient être considérées comme des aspects de l'intelligence non descriptibles « de manière si précise qu'une machine peut les simuler ».

Post, qui était d'abord arrivé à la même conclusion à la suite de ses propres travaux sur l'indécidabilité, s'est ensuite montré d'un autre avis. Son raisonnement se base sur le fait qu'une fois que la pensée est explicite, elle peut être simulée par une machine[475]. Or cela vaut également pour *tout argument contre* le fait que la pensée puisse être simulée par une machine : une fois que l'argument est formulé, il peut être simulé par une machine. De la même manière, il est possible de retourner le procédé utilisé par Gödel – la construction d'une proposition auto-référente, appelée diagonalisation, et la mécanisation de la construction de cette proposition – pour défaire les arguments antimécanistes :

> La découverte essentielle de Gödel a été la mécanisation de la diagonalisation, de sorte que son argument antimécaniste se heurte à un dilemme : ou bien il est non effectif, ou bien il est incapable de montrer que celui qui le formule n'est pas une machine. En d'autres termes, dès le moment où nous appliquons l'argument gödelien à une machine quelconque pour en conclure que nous ne pouvons pas être modélisés par cette machine, l'argument en question pourrait être modélisé par une machine[476].

On reconnaît ici l'impossibilité de fournir un contre-argument à l'hypothèse formulée à l'occasion du séminaire de Dartmouth (voir section 1.1.10. Le paradoxe du contre-argument). Si le contre-exemple est bien descriptible, il peut être simulé par une machine, et ce n'est pas un contre-exemple. Si, à l'inverse, le contre-exemple est peu descriptible, il ne sera pas reçu comme un contre-exemple sérieux. On lui reprochera d'être flou et mal défini. En conséquence, les théorèmes de Gödel apportent une restriction quant aux possibilités du formalisme mais ne permettent pas pour autant d'apporter un argument définitif en faveur de l'intuition. À partir de considérations similaires sur l'indécidabilité, Gödel conclut de l'impossible mécanisation de l'esprit, alors qu'au contraire Post en tire un moyen de défaire tout argument explicite contre la mécanisation de l'esprit.

475 « Post, par exemple, après avoir adopté une position analogue à celle de Gödel, l'avait abandonnée. Une fois découvert son premier résultat d'indécidabilité, en 1920, il en avait conclut la fausseté du mécanisme. Puis il avait été amené à poser un « axiome de réductibilité pour les opérations finies », équivalent à la thèse de Church, où il ne voyait qu'une hypothèse de travail demandant à être justifiée : si le processus créatif commence dans l'inconscient, il lui est essentiel de pouvoir devenir conscient, explicite, effectif. Toute procédure humaine répondant à ces conditions peut alors être simulée par une machine. » Michel Bourdeau, *op. cit.*, p. 111.

476 Michel Bourdeau, *op. cit.*, p. 110-111.

À la suite de Gödel, Lucas[477], puis Penrose[478], les théorèmes d'incomplétude ont été considérés comme des preuves qu'un système formel ne peut décrire l'ensemble de la pensée – établissant ainsi la nécessité de l'intuition. Mais la question reste controversée, les arguments de Lucas et Penrose ont été critiqués par Putnam, Benacerraf, Quine et d'autres[479]. Sans entrer dans les détails de la controverse, nous retenons que les limites du formalisme ne semblent pas permettre pas d'extrapoler avec certitude de la nécessité de l'intuition. Autrement dit, formaliser la limite du formalisme ne permet pas de formaliser ce qui pourrait se situer hors de ces limites.

Si la portée des théorèmes d'incomplétude est reconnue dans le champ des mathématiques, la pertinence de leur extrapolation à l'intelligence artificielle fait débat. Les théorèmes d'incomplétude laissent planer un doute sur la pertinence des hypothèses du projet d'intelligence artificielle, mais ils n'ont pas permis de produire de réfutation incontestable – et peut-être en sont-ils incapables. Ils laissent penser qu'il y a une impossibilité à remplacer le travail de l'intuition par une procédure explicite, une machine, mais sans qu'il soit pour autant possible de le prouver. Autrement dit, il semble impossible d'évaluer le rôle de l'intuition de manière purement formelle, sans recourir à l'intuition. Tout se passe comme si nous ne pouvions faire mieux que de partager l'intuition de la nécessité de l'intuition, mais sans que cette intuition ne puisse trouver de confirmation par le raisonnement explicite. En cela, la nécessité de l'intuition ressemble aux propositions indécidables dans la mesure où elle est orpheline d'un raisonnement explicite permettant de la fonder. Si l'intuition correspond à la part de la pensée qui échappe à la description par un système formel, son existence même doit rester une proposition indécidable. Si elle était décidable, quel que soit le résultat de la décision, cela contredirait sa définition : soit il est prouvé que l'intuition n'existe pas, et alors il serait possible de réduire l'ensemble de la pensée à un système formel ; soit il est prouvé qu'elle existe, mais la preuve de son existence reviendrait à l'inclure dans un système formel décrivant la pensée, ce qui contredirait sa définition.

En d'autres termes, dès qu'un aspect de l'intuition reçoit une preuve formelle, cet aspect ne relève plus de l'intuition, puisqu'il a été prouvé. L'intuition devrait demeurer, par définition, hors du domaine de la preuve. **S'il semble vain de vouloir se passer l'intuition, il est tout aussi vain de vouloir fonder sa nécessité. Elle résiste aux tentatives d'élimination, mais ne se prête pas pour autant à une entreprise de fondation.** Nous ne pourrions rien faire de

477 John Lucas, « Minds, Machines, and Gödel », *Philosophy,* n° 36, 1961, p. 112-127
478 Roger Penrose, *Emperor's New Mind*, Oxford, Oxford University Press, 1989.
479 Voir Stanislaw Krajewski, « On Gödel Theorem and Mechanism : Inconsistency or Unsoundness is Unavoidable in Any Attempt to 'Out-Gödel' the Mechanist », *Fundamenta Informatica,* n°81, 2007, p. 1-9.

mieux que, à l'instar de Hilbert, séparer ce qui tient au régime des preuves et des démonstrations (pensée formelle) et ce qui relève des axiomes et de la métamathématique (intuition).

2.5.4. Complémentarité et division du travail

La pensée symbolique est une prothèse qui permet de remédier à la fragilité de l'intuition. Cependant, l'efficacité de la prothèse n'implique pas pour autant qu'on puisse se débarasser de l'intuition. Pour filer la métaphore visuelle, discréditer la pensée formelle reviendrait à se priver de lunettes alors que l'expérience a montré que notre vue est très mauvaise. Mais croire que l'on peut se passer de l'intuition et réduire la pensée à son formalisme reviendrait à s'arracher les yeux sous prétexte qu'on a d'excellentes lunettes. Si l'intuition est devenue *persona non grata* dans le domaine de l'intelligence artificielle, elle a fait au contraire l'objet d'une véritable « réhabilitation » en mathématiques, « dans le champ d'une discipline tenue pour le modèle insurpassé de la rigueur[480] », en particulier après l'effort de formalisation mené à la fin du dix-neuvième et au début du vingtième siècle. Pour Michel Bourdeau, cela « suffit déjà à montrer ce que peuvent avoir d'outré les accusations de ceux qui ne veulent voir dans l'intuition que l'antichambre de la déraison[481]. » En d'autres termes, l'importance de l'intuition en mathématiques, la discipline où le formalisme est poussé à son comble, montre que les réticences qu'on peut avoir à son égard ne suffisent pas à motiver son élimination et laissent penser que le formalisme et l'intuition peuvent faire bon ménage – voire ne peuvent *que* faire bon ménage, l'un ne fonctionnant pas sans l'autre.

Nous avons déjà mentionné comment l'intuition, si elle est exclue des mathématiques par le programme de Hilbert, garde un rôle de premier plan puisqu'elle est la seule à pouvoir s'occuper de métamathématiques (voir section 2.4.2. Le programme de Hilbert). Le projet formaliste aboutit à une tentative de *division du travail* entre la pensée formelle, qui s'occupe de mathématiques et l'intuition, qui s'occupe de métamathématiques.

> L'originalité majeure du programme de Hilbert réside dans la séparation rigoureuse et systématique établie entre deux points de vue, qui a pour effet de rendre équivoque la notion de raisonnement. Selon qu'on raisonne dans le système ou sur lui, dans la mathématique

480 Michel Bourdeau, *op. cit.,* p. 105.
481 *Ibid.*

> ou dans la métamathématique, démontrer sera tantôt déduire selon les règles du calcul, tantôt montrer par des procédés intuitifs[482].

Michel Bourdeau compare cette division du travail à une séparation entre production artisanale et production manufacturée. La mathématique formelle « sert non à décrire la production artisanale des théorèmes, mais à donner de cette dernière des reproductions manufacturées, indispensables au travail, à nouveau artisanal, du métamathématicien[483]. » Nous retrouvons la distinction qu'opérait Turing en 1938, entre l'« oracle » de l'intuition, fabriquant des « jugements spontanés », et les opérations de l'*ingenuity,* fonctionnant comme une machine. En ce sens, la distinction entre intelligence artificielle et naturelle peut se faire bien en amont de la fabrication de machines. Dans la tête du mathématicien, il y aurait alternance, voire cohabitation, entre l'artisanat métamathématique (naturel) et la production standardisée mathématique (artificielle). Lorsqu'un propos, rigoureusement formalisé, a une portée métamathématique, celui qui l'écrit et ceux qui le liront font cohabiter l'*ingenuity* (vérifier la conformité aux règles) et l'intuition (percevoir les enjeux métamathématiques).

Une fois produit, le travail artisanal du métamathématicien pourra être formalisé et constituer de nouvelles reproductions manufacturées au service des autres artisans métamathématiciens. La métaphore peut s'appliquer à d'autres domaines que les mathématiques. Des formes préexistantes plus ou moins élaborées, achevées (théorèmes préexistants, textes travaillés, tableaux) ou inachevées (images, mots, sons, bribes de conversation, fragments d'idées…) évoluent et coalescent progressivement jusqu'à un degré variable de cohérence, dont la formalisation mathématique est un extrême. La forme ainsi acquise s'agrége ensuite à l'ensemble des 'pensées toutes faites' : clichés, arguments préconstitués, théorèmes manufacturés – qui constituent une boîte à outils à disposition de la pensée pour produire de nouvelles formes.

Si l'intuition est au théorème constitué ce que l'artisanat est au produit manufacturé, alors elle le précède. On doit à Brouwer d'avoir, tout à l'inverse de Hilbert, fondé les mathématiques sur l'expérience de l'intuition[484]. Pour Hilbert, c'est la rigueur formelle d'un énoncé qui permet de vérifier un énoncé, tandis que pour Brouwer, c'est la possibilité d'en faire l'expérience. Ce qui décide de la validité d'un énoncé, c'est qu'un esprit puisse le construire.

---

482 *Ibid*, p. 29.
483 *Ibid*, p. 36.
484 Mark Van Atten, *On Brouwer,* Wadsworth Philosophers Series, 2004. Voir aussi Jean Largeault, *Intuition et intuitionnisme*, Paris, Vrin, 1993. Les textes de Brouwer ont été traduits et publiés dans l'ouvrage de Jean Largeault (ed.), *Intuitionnisme et théorie de la démonstration,* Paris, Vrin, 1992.

Hilbert conçoit les mathématiques comme un jeu de langage hors du temps. Brouwer, au contraire, les fait dépendre de l'intuition, et donc du temps. Le langage formel ne fait qu'accompagner les mathématiques, tout comme une carte météo cherche à « suivre » un phénomène météorologique[485]. Ce qui ne veut pas dire que les mathématiques peuvent se passer de travail : il faut un effort spécifique pour en effectuer les constructions mentales. En mettant l'accent sur la construction, l'intuitionnisme invite à un point de vue particulier sur l'informatique. Cette dernière permettant de simuler les étapes qu'un esprit devrait effectuer pour construire un objet mathématique, peut-on dire qu'elle revient à simuler l'intuition du mathématicien ?

2.5.5. La bête à deux dos

La distinction entre pensée aveugle et intuition nous aide à en apprendre un peu plus au sujet du processus de pensée. Mais il ne faut pas prendre ce qui n'est une distinction instructive pour un conflit dans lequel il faudrait prendre parti. Trop souvent, l'éloge de la pensée aveugle est confondu avec celui de la rigueur. On veut prendre parti pour les règles contre l'inconstance de l'intuition. **Mais négliger l'intuition revient à manquer de rigueur d'une autre manière : c'est porter un discours sans prendre en compte les hypothèses et les vérités premières qui le soutiennent – en faisant comme si elles étaient évidentes pour tous –** ce qui peut être un artifice rhétorique, voire de la mauvaise foi, mais certainement pas une manière juste d'exposer ses idées. Pour penser correctement, il faut tenir compte des principes sans s'effaroucher de leur indécidabilité. En face, les « partisans » de l'intuition se veulent les chantres de l'authentique et du naturel. Mais, au risque de passer pour inexistentes, leurs intuitions doivent laisser au moins une trace, au mieux une forme achevée, de laquelle on attend un minimum de consistance pour qu'elle soit communicable, sinon l'affaire est vaine. On ne saurait s'affranchir de règles de précision et de lisibilité, voire de vérifiabilité. Les deux positions se retrouvent dans un élan commun, le même désir suspect de pureté : pureté de la règle, de la certitude partagée, de la forme achevée qu'on peut reproduire et revérifier ; ou pureté de l'immédiat, de la nature, de l'authentique. Mais il est vain de vouloir choisir entre Papa (la rigueur de la forme) et Maman (l'authenticité de la nature). Il faut les deux, et l'obscénité de leur mélange, pour donner

485 « Formal language accompanies mathematics as the weather-map accompanies the atmospheric processes », Brouwer, cité par Mark Van Atten, *On Brouwer, op. cit.*, p. 451.

naissance (Maman) à une pensée qui ait quelque *consistance* (Papa). Prendre parti pour l'un ou pour l'autre sans nuances, c'est se demander si la jambe gauche vaut mieux que la droite. Jean Lassègue remarque ainsi, à propos de la distinction effectuée en 1938 par Turing entre *intuition* et *ingenuity,* que « la frontière entre les deux facultés n'est cependant pas fixée une fois pour toutes », en effet

> les productions non-constructives de l'intuition peuvent devenir constructives grâce au travail logique et algorithmique émanant de l'ingéniosité. Les rapports entre l'intuition et l'ingéniosité sont donc de nature essentiellement dynamique et c'est bien en cela qu'il s'agit d'activités de pensée et non pas seulement de facultés dont les domaines de validité seraient conçus de façon figée[486].

Plus précisément, les deux facultés paraissent « empiéter l'une sur l'autre : »

> l'ingéniosité vient se substituer à l'intuition en lui donnant une expression par le biais d'un algorithme mais inversement, la constitution du concept de machine de Turing, concept 'ingénieux' par excellence, est le résultat d'un acte d'intuition complètement original. Il faut donc envisager les rapports entre les deux facultés sur un mode dynamique : chaque fois que l'on trouve un algorithme susceptible d'être exécuté par une machine de Turing, c'est-à-dire chaque fois que l'on étend le domaine de validité de l'ingéniosité, on retrouve cependant en même temps la trace de l'intuition qui a présidé à la naissance du concept de machine de Turing487.

La distinction entre ingénuité et intuition a pour intérêt principal de nous permettre de mieux appréhender ce qu'est l'intuition, cette dernière étant, des deux dos de la bête, le moins connu, bien que la philosophie n'ait jamais cessé de l'étudier. Il reste toujours plus facile d'enseigner et de transmettre les règles simples et les définitions claires des mathématiques plutôt que la pratique du doute et l'expérience de la précarité des vérités premières. Pour conclure notre parcours sur l'intuition, nous allons maintenant reprendre l'ensemble des traits constitutifs que cela nous a permis d'identifier et les récapituler dans une définition.

486 Jean Lassègue « L'évolution du constructivisme turingien : de la logique à la morphogenèse », *op. cit.* p. 112.
487 *Ibid,* p. 113.

2.6. Définir l'intuition

2.6.1. Principes de l'intuition

L'intuition est difficile à approcher et à définir, bien qu'elle soit si familière. Il est possible de lui donner de la consistance en passant par la comparaison avec les organes de la perception, au premier rang desquels la vision, mais aussi l'odorat, l'ouïe ou le *gut-feeling* ; ou par contraste avec la « pensée aveugle », ce qui est une manière d'étudier l'intuition lorsqu'elle est mise en suspens, – à l'instar des physiologistes qui étudient un organe en procédant à son ablation.

La définition la plus simple qui puisse en être donnée est la suivante : l'intuition est ce qui nous permet de comprendre. Elle est **ce qui perçoit le sens d'une forme** quelconque : un énoncé une image, un geste, une situation. La notion de sens pouvant s'entendre de plusieurs manières, à la fois comme ce à quoi renvoie la forme en tant que signe (sa dénotation), mais aussi ce qu'elle vise (la conclusion d'un texte, le résultat d'une démonstration), ce qu'elle sert (sa finalité), ou encore les principes dont elle découle (« vérités premières » pour un texte, axiomes pour un théorème ou archétypes pour une image).

Lorsqu'elle perçoit la visée d'une forme, l'intuition se montre capable d'anticipation. Elle joue le rôle d'un « oracle », selon le mot de Turing, qui donne une idée de ce que sera le résultat d'un problème avant que nous n'ayons commencé à en étudier les détails. Dans le meilleur des cas, elle guide notre raisonnement vers ce résultat qu'elle pressent. Elle nous en donne une idée vague, qui se précise à mesure que nous effectuons ce qu'elle nous invite à faire – comme dans une enquête de Sherlock Holmes ou du commissaire Maigret[488]. C'est également l'intuition qui nous signale lorsque le résultat est atteint, par la perception d'une adéquation (un sentiment d'« eurêka ») qui peut néanmois être trompeuse.

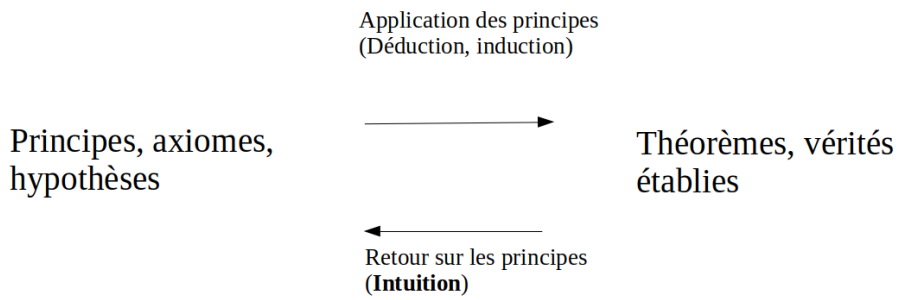
Lorsqu'elle perçoit les principes d'une forme, l'intuition fonctionne dans un **mouvement inverse** à l'intelligence habituelle. Au lieu de « suivre » un raisonnement, elle en

488 Par exemple, dans *Les vacances de Maigret*, celui-ci sent immédiatement, sans que cela soit rapporté explicitement, qui est le coupable (le docteur Bellamy). Cela conduit le commissaire à une série de démarches erratiques incompréhensibles pour ses pairs, voire impolies, d'autant plus qu'il ne cesse de répéter qu'il est en vacances et que l'affaire ne le regarde pas. Jusqu'à ce que l'assassin, voyant que la prémonition de Maigret l'amène à trouver les indices qui la confirment, finisse par se livrer de lui-même. Georges Simenon, *Les vacances de Maigret*, Paris, Le Livre de Poche, 2001.

« remonte » le cours pour en interroger les hypothèses. Son mouvement de récursion l'amène, de « raison » en « raison », à la part du raisonnement qui ne dépend plus d'aucune autre « raison » et ne se justifie pas – autrement dit à ce qui est *originaire,* ce qui, selon les mots de Silesius commentés par Heidegger, est, comme la rose, *sans pourquoi*[489]. C'est la définition qu'en donne David Rabouin :

> L'intuition, en ce sens, c'est plutôt l'accès à ce qui est premier, à partir de quoi on procède ensuite (en dé-duisant) : vérités primitives, principes, fondements, axiomes, « idées claires et distinctes » ou données brutes des sens, selon le goût philosophique de chacun[490].

Ce que nous pouvons illustrer par le schéma suivant.



Aussi peut-on dire que l'intuition s'intéresse à ce qui, dans la pensée, est le plus élémentaire, le plus simple, mais aussi le plus difficile à rendre explicite. Tout ce que, selon le mot de Saint-Augustin à propos du temps[491], nous connaissons sans être en mesure de bien le définir : le temps, la matière, la vie, et autant de notions qui sont l'affaire de la philosophie. Ces concepts sont « simples » au sens où ils ne sont pas « composés ». Ils sont élémentaires, on ne peut renvoyer qu'à eux-mêmes pour les définir. Mais ils sont difficiles à concevoir, bien qu'ils ne semblent poser aucune difficulté lorsqu'ils sont intégrés à des énoncés composés. Il est aisé de comprendre la phrase « nous disposons de peu de temps », ou bien « la vie se sert de la matière pour croître ». Il est autrement plus difficile de comprendre les notions de « temps », de « vie » ou de « matière ». Tout comme la « pensée aveugle », mais d'une manière très différente, l'intuition nous permet d'utiliser des notions que nous comprenons mal. Nous savons ou nous

489 Martin Heidegger, *Le principe de raison,* Paris, Gallimard, 1962, p. 102-109.
490 David Rabouin, *op. cit.*
491 « Qu'est-ce donc que le temps ? Si personne ne me le demande, je le sais ; mais si on me le demande, je ne le sais plus. » Augustin d'Hippone, *Les Confessions,* Paris, Garnier Flammarion, 1964, Livre XI, Chapitre XIV.

« sentons » ce que signifie le temps, sans le « voir » tout à fait. De ce point de vue, l'intuition est loin de la « représentation si facile et si distincte qu'il ne subsiste aucun doute sur ce que l'on y comprend [492] ». Au contraire, elle avance à tâtons, dans une obscurité qui n'est toutefois pas la même que celle où règne la « pensée aveugle ».

La notion d'intuition, qui nous permet d'appréhender les « vérités primitives » sans les comprendre tout à fait, semble être de même nature : la notion d'intuition est aussi une « vérité primitive » sur laquelle repose notre conception de la pensée sans que nous soyons capables de la rendre explicite de manière pleinement satisfaisante. Il en va de même pour les notions voisines que sont le sens et la compréhension. Il est facile de restituer le sens d'une phrase comme « je comprends ce que l'examinateur attend de moi », mais autrement plus difficile de rendre compte de ce que veut dire « je comprends ».

À ces difficultés que posent l'intuition s'ajoute sa variabilité. La trace d'une intuition, même lorsqu'elle a été soigneusement mise en forme, n'apporte aucune certitude de sa réitération : j'ai peu de garanties que le texte que j'écris aujourd'hui fera sens pour moi demain, et encore moins qu'il fera sens pour les autres. En raffinant la forme (ici, le texte), je peux augmenter les chances d'une transmission réussie, mais je n'atteindrai pas le degré de certitude fourni par la « pensée aveugle ». Une fois mis au point, un algorithme comme le PGCD peut être effectué n'importe quand et par n'importe qui, il donnera toujours le même résultat. À l'inverse, le cheminement que propose Descartes dans les *Méditations cartésiennes* pour « faire sentir », donner l'intuition du *cogito* à son lecteur, n'aboutit pas à chaque fois. Si je l'enseigne, il me faudra probablement m'y reprendre à plusieurs fois pour que chaque élève « comprenne ». Et il y a fort à parier que, l'un d'eux retombant sur le texte quelques années après, ne garde qu'un vague souvenir de la classe et se retrouve incapable de sentir à nouveau l'intuition du *cogito.* L'intuition s'effectue en un temps et un lieu singuliers, alors que la forme achevée peut se reproduire n'importe où et n'importe quand – comme si elle s'exprimait depuis un point de vue de nulle part. L'œuvre se trouve comme émancipée du temps et du lieu, et peut être répétée à l'envi, par contre son interprétation se fait toujours **au présent.** Elle doit, comme le feu, être constamment ravivée. Voilà pourquoi elle mobilise l'attention, alors que la « pensée aveugle » peut s'effectuer sans y penser. Alors que les déductions peuvent être abstraites du temps et accumulées, l'intuition ne se laisse pas stocker. Ce que j'ai compris auparavant m'aide à comprendre ce que je lis aujourd'hui, l'expérience accumulée m'apporte de l'aisance, mais elle ne me dispense pas de l'effort : je dois, à chaque fois, recommencer le geste. L'exercice de la

492 René Descartes, *Règles pour la direction de l'esprit,* Paris, Le Livre de Poche, 2002, p. 85.

« pensée aveugle » peut se répéter à l'identique alors que l'interprétation, la transmission de l'intuition, exige toujours des reformulations et de nouvelles traductions. Faire commerce avec les principes et vérités premières s'effectue toujours au présent.

Voici ce qui, pour reprendre les termes de Turing, serait le « réel » de la pensée, ce qui « reste » une fois enlevées toutes les peaux de l'oignon, c'est-à-dire toutes les fonctions mécaniques, automatiques, que l'on peut effectuer sans y penser. La seule chose qu'on ne peut effectuer sans y penser, c'est ce « y penser ». On peut probablement mécaniser toutes les fonctions de l'esprit, mais pas le fait de penser à ce que l'on fait lorsqu'on effectue une de ces fonctions. La compréhension ne peut se mécaniser car elle mobilise des principes qui, bien que simples, ne peuvent être rendus explicites. Leur intuition peut se transmettre mais non s'abstraire du temps et de l'espace pour être intégrée à un mécanisme.

Ce « y penser » est à la fois le plus proche et le plus difficile à percevoir, puisqu'il s'efface devant ce qu'il donne à comprendre. Quand j'effectue une fonction en « y pensant », c'est la fonction que j'ai à l'esprit, fonction qui paraît mécanisable. Partout où je porte mes yeux, toutes les fonctions que j'effectue en y pensant, me semblent mécanisable. Mais ce serait oublier *ce qui me permet de le voir,* cet œil de la pensée auquel ressemble l'intuition. Tout est mécanisable, peut-être, sauf l'œil pour qui tout est mécanisable.

2.6.2. L'intuition de l'instant

La plupart de nos considérations sur l'expérience a trait à ce qui, de l'expérience, *se répète*[493]. J'interprète une situation grâce à ce qu'elle me rappelle de situations similaires. Si un fait nouveau vient enrichir mon expérience, j'en « tire la leçon » pour l'avenir. En d'autres termes, je prête une grande attention à ce qui, d'une situation à l'autre, constitue une série de caractéristiques similaires me permettant d'en faire des cas paradigmatiques pour interpréter le passé ou anticiper sur l'avenir. D'une situation à l'autre, je tisse patiemment une toile d'araignée de similarités dans laquelle capturer les événements et en tirer le meilleur parti.

Il faut une naissance, un décès, ou tout autre événement suffisamment perturbant, pour que l'irréductible nouveauté de ce qui a lieu me contraigne à tourner mon regard, une fois n'est pas coutume, vers *ce qui ne s'y répète pas.* Naissances et décès ont beau ressembler aux autres

493 Les considérations qui suivent, et le titre de la section, sont indirectement inspirés de l'ouvrage de Gaston Bachelard, *L'intuition de l'instant,* Paris, Le livre de poche, 1994.

naissances et décès, j'ai le sentiment qu'il *se passe quelque chose*. Je vis un instant d'ivresse ou de panique où je réalise que la totalité de ce qui se trame me rend incapable de constituer une impression de ce qui a lieu. Ici et maintenant, les parties du monde forment une conjonction qu'elles ne referont plus jamais. Le tissu de caractéristiques connues que je me suis patiemment constitué afin de capturer les similarités entre événements est pris en défaut et ne fait que souligner la singularité de ce qui se trame. Cette *différence* ne trouve en moi aucun support, aucun moyen ou media, pour que je m'en forge une représentation. Bien que je participe à cette rencontre inédite entre parties du monde, je n'ai que confusément accès à sa nouveauté.

Le sentiment de répétition peut alors s'inverser : ce qui se répète d'une situation à l'autre, ce qui s'est toujours répété, c'est leur irréductible nouveauté. C'est un point commun de toutes les situations que j'ai vécu, qui traverse de part en part l'ensemble de mon expérience : chacune d'entre elles n'a eu lieu qu'une seule fois. Elles ont toujours été radicalement nouvelles. Ma naissance ou mes premiers pas n'ont été perçus que sous l'angle de cette stupéfaction primordiale qui se passe de représentation. Depuis ces premières expériences, chaque situation que je vis est tout aussi *première,* mais l'étonnement s'est émoussé.

L'étonnement est toujours là, mais parce qu'il est toujours là, parce qu'il est si commun à toutes les expériences, je ne le vois plus. *Je ne m'en étonne plus.* Bien que j'aie depuis le premier jour un sentiment vif de la singularité des situations, je préfère ne m'entretenir que de leurs points communs et des répétitions. Il faut l'étrangeté d'un *déjà-vu* pour restituer cet étonnement primordial : en me troublant par le sentiment que *tout se répète à l'identique, jusque dans les moindres détails,* le déjà-vu manifeste qu'une situation qui se répéterait *exactement* est anormale, d'une anormalité plus grande que le fait que chaque situation diffère légèrement des autres. Ainsi, pour m'apparaître comme *normale* une situation ne doit *jamais avoir été vue*. Contrairement à ce que mes ruminations conscientes pourraient laisser penser, je suis profondément habitué à que ce que je vive soit *exceptionnel.*

Il existe un hiatus formidable entre ce sentiment de l'exception – si quotidien que je l'oublie – et les représentations dont je m'entretiens. L'incessant théatre de mon « for intérieur » ne s'intéresse qu'aux *régularités* pour en tirer des *règles utiles,* mais presque jamais à l'exception. Tout se passe comme si une certaine *paresse d'esprit* m'amenait spontanément à ne penser qu'aux répétitions et à éviter soigneusement les singularités.

Cette paresse peut s'attribuer à l'effroi que suscite la considération de la singularité de l'instant. Au fait qu'il ne répète rien et ne se répètera jamais, s'ajoute le fait qu'il ne me laisse jamais le temps de l'appréhender pleinement. A l'instant où je tourne mon regard vers lui, l'instant a déjà traversé la passoire de mon esprit. Il ne s'arrête jamais pour que j'aie le loisir de

le contempler[494]. De mon point de vue, le présent n'a donc jamais *pleinement lieu.* Je n'en prend conscience que *déjà perdu.* Non seulement la singularité d'une situation ne se répète pas, mais en plus elle ne s'est jamais pleinement donnée à voir. En ce sens l'intuition de l'instant est moins l'intuition d'une présence que l'intuition d'une perte. La prise en compte, dans l'instant, de ce qui ne s'y répète pas, est contemplation d'une *perte incessante*. C'est une chute vers l'absence. Un vertige : il ne s'agit pas de se faire peur en regardant le vide depuis l'abri d'un balcon, mais de réaliser que *nous sommes toujours en train de chuter.* L'intuition, en ce sens, est l'expérience d'une singularité, d'une exception, et donc toujours d'une perte, que chaque expérience répète silencieusement.

2.6.3. Concevoir une forme

Revenons à la part de l'intuition qui a trait à la perception de formes répétables. L'intuition y est également singulière (à renouveler selon le temps et le lieu), inconstante, personnelle. Elle ne s'objective pas, elle ne peut que se transmettre. Elle est, comme l'indiquent les comparaisons avec la lumière, spontanée, « naturelle », et s'oppose à la forme constituée qui est fixe, artificielle. Quand l'œuvre achevée (la forme constituée) fait l'objet d'une maintenance, d'une conservation telle quelle, son interprétation doit être renouvelée en tenant compte de la singularité du moment et du lieu. Pour saisir ce que l'intuition effectue, le mot de *conception* semble plus adapté que celui de *compréhension*. Lorsque je comprends une idée, je la *conçois*, c'est-à-dire que je la forme dans mon esprit, dans un mouvemement qui est toujours une première fois, une naissance – qu'il s'agisse d'une idée nouvelle ou d'une idée bien connue.

D'après Michel Foucault, il faut combiner deux niveaux de lecture pour bien lire les *Méditations* de Descartes. Les *Méditations* se lisent comme « un enchaînement systématique de propositions, moments de pure déduction[495] » qu'il appelle « la trame démonstrative ». En tant qu'elles sont des *méditations*, elles doivent également se lire comme « un ensemble de modifications formant exercice, que chaque lecteur doit effectuer, par lesquelles chaque lecteur doit être affecté, s'il veut être à son tour le sujet énonçant, pour son propre compte, cette

494 Et les captures partielles que je peux en faire (notes, photos, vidéos) n'y changent rien, au contraire, elles redoublent la perte en faisant signe vers ce qu'elles n'ont pu en capturer. Roland Barthes, *La chambre claire, Note sur la photographie*, Paris, Cahiers du cinéma Gallimard, 1980.

495 Michel Foucault, « Mon corps, ce papier, ce feu », *Dits et écrits, 1954-1988*, tome 2, Paris, Gallimard, 1994, p. 258.

vérité[496] ». Foucault appelle ce deuxième mode de lecture la « trame ascétique », le mot « ascétique » renvoyant ici à l'*askesis*, l'exercice. Nous retrouvons une manière d'opposer la pensée aveugle et l'intuition, à la différence que l'intuition ne renvoie pas à une perception passive mais à une pratique active, une opération d'effectuation du sens du texte.

Il est tentant de renvoyer l'opposition entre pensée aveugle et intuition à la distinction entre contenant et contenu. La pensée aveugle serait l'affaire du contenant, de la forme extérieure. L'intuition, quant à elle, s'occuperait de percevoir le contenu, le sens « contenu » par la forme. Mais la notion de « contenu » laisse penser qu'il y a une chose qui attend, tapie au sein du contenant, qu'on la perçoive. La notion de conception permet d'opérer un décalage : **le sens n'attend pas qu'on le percoive, il attend qu'on l'effectue**. En ce sens, **il ne préexiste pas à son interprétation**. Le concert n'est pas « contenu » dans la partition, il ne préexiste pas à son « interprétation » par le pianiste. Si le sens était une « chose », « contenue » par la forme, il pourrait être rendu explicite. Il reste implicite car il est ce que la forme m'invite à effectuer – qui se traduit aussitôt par de nouvelles formes. Quand on « suit » une intuition, ce que l'on découvre après-coup n'est pas « ce qui était contenu », mais ce qui était en germe. On réalise, dans les deux sens du terme, *ce que nous a fait faire une forme,* ce qu'elle nous a poussé à formaliser. La marche de l'intuition n'est pas hypothético-déductive au sens où la « déduction » vient recouvrir et confirmer une hypothèse. Il y a déploiement d'un germe, genèse. L'intuition ne donne pas accès à un « arrière-monde » d'idées constituées, flottant dans une limbe immatérielle, qui attendent d'être perçues. Les idées ne préexistent pas à leur conception, pas plus que je n'existais avant que mes parents ne me conçoivent. L'intuition est le « naturel » des idées au sens de leur naissance et de leur croissance *immanentes.*

Aussi, la distinction entre pensée formelle et intuition n'est pas aussi rigide qu'il n'y paraît. Si on adopte une acceptation large de la « pensée formelle » qui inclut toute pensée s'appuyant sur des formes plus ou moins autonomes par rapport au penseur (images, langage, sons…), on voit mal comment attribuer une consistance à un deuxième terme de l'opposition : l'idée d'une « pensée sans forme » qui précèderait toute forme. Elle aurait l'avantage d'être « pure » de tout artifice, radicalement « originaire » et « naturelle ». Elle satisferait à un fantasme suspect de pureté et d'immédiateté, mais on voit mal comment se la représenter et se convaincre de son existence. Autrement dit le « réel » de la pensée ne se trouve pas au cœur de l'oignon mais *dans chaque peau,* au moment de sa conception. Une fois la peau conçue, elle

496 *Ibid,* p. 264.

peut être séparée de l'oignon : la forme conçue par l'esprit peut être inscrite sur un support et prendre une relative autonomie.

L'opposition entre pensée formelle et intuition doit être précisée en une distinction entre *pensée formalisée* et *processus de formalisation*, pensée constituée et conception de la pensée. L'intuition ne se place pas avant la forme, elle s'exerce dans la conception de la forme. En l'effectuant (trame ascétique), en la conçevant selon la singularité de son esprit, de son temps, de son lieu, notre esprit en produit sa propre version. Concevoir une idée, c'est la répliquer, c'est l'interpréter, comme on dit d'un musicien qu'il interprète une partition sauf qu'il n'y a pas, en l'occurence, de partition, d'idée préexistante, d'arrière-monde des idées. Comme les virus, les formes circulent et se répliquent d'esprit en esprit. Chaque esprit, en les hébergeant, leur fournit l'occasion de muter (d'être réinterprétée) à la faveur d'une rencontre avec un mélange singulier d'autres formes. Nous interprétons une nouvelle forme en fonction de la collection de formes que notre mémoire entretient déjà. Ainsi, la *conception* d'une idée est aussi une affaire d'*engendrement* : engendrement de l'idée pour mon propre compte quand je fais, par exemple, l'expérience du *cogito* cartésien ; mais aussi engendrement d'autres formes, par rencontre et hybridation avec les autres autres formes que j'héberge. Bien comprendre une idée, la concevoir pour soi, c'est sentir émerger comme elle fait résonner d'autres formes connues et en engendre de nouvelles : commentaires, questions, reflexions…

Autre conséquence, il n'y a pas de vie des idées sans humains pour l'héberger. Les formes ont besoin de nous comme les virus ont besoin de cellules vivantes pour se répliquer et se perpétuer. En filant la métaphore, les pressentiments qu'offrent l'intuition peuvent être comparés à la période d'incubation d'un virus. Lorsque je pressens une solution, c'est que la forme est déjà là, mon esprit a commencé à la concevoir sans qu'elle ne soit encore manifeste et il faudra, pour cela, un effort intense et parfois douloureux. Autre point commun entre le virus et l'idée : celui qui répète une forme sans en voir le sens serait comme un porteur asymptomatique. Il véhicule une idée (un virus) sans en ressentir les effets (les symptômes), et ses chances de contaminer les autres sont bien plus faibles. Toutefois, la comparaison avec le virus a ses limites. L'expérience de la maladie est douloureuse et me fragilise tandis qu'avoir une idée et la formaliser reste, malgré les difficultés, une expérience agréable qui me donne le sentiment de me réaliser.

Henri Poincaré décrit ainsi comment les idées bouillonnent *spontanément* dans son esprit. Elles semblent se chercher jusqu'à trouver le moyen de s'associer.

> Depuis quinze jours, je m'efforçais de démontrer qu'il ne pouvait exister aucune fonction analogue à ce que j'ai appelé depuis les fonctions fuchsiennes ; j'étais alors fort ignorant ; tous les jours, je m'asseyais à ma table de travail, j'y passais une heure ou deux, j'essayais un grand nombre de combinaisons et je n'arrivais à aucun résultat. Un soir, je pris du café noir contrairement à mon habitude ; je ne pus m'endormir ; les idées surgissaient en foule ; je les sentais comme se heurter, jusqu'à ce que deux d'entre elles s'accrochassent pour ainsi dire pour former une combinaison stable. Le matin, j'avais établi l'existence d'une classe de fonctions fuchsiennes, celles qui dérivent de la série hypergéométrique ; je n'eus plus qu'à rédiger les résultats, ce qui ne me prit que quelques heures[497].

On notera l'ambigüité de la formule « j'avais établi » : *qui* a établi l'existence d'une classe de fonctions fuchsiennes ? Pendant quinze jours Poincaré a fait de grands efforts, mais lors de la dernière soirée, il n'a rien fait d'autre que prendre un café. Une grande partie de l'effort consiste à sentir, et surtout à consentir à ce qui est perçu par l'intuition. Il n'a été que l'hôte passif de cette sarabande des idées cherchant d'elles-mêmes « une combinaison stable ». Le travail de l'esprit exige une part d'assiduité, de volonté, de pugnacité, qui se mêle à une part involontaire où l'activité ressemble à la croissance du végétal. Quelque chose (l'idée d'une classe de fonctions fuschiennes) a cru en Poincaré, l'a pris comme substrat pour se déployer. Comme le fruit d'un arbre, l'idée a poussé « de » son esprit. On peut dire qu'il a « fait » un théorème comme on dit d'un enfant qu'il « fait ses dents[498] ».

2.6.4. Nature de la culture

L'intuition ne s'oppose pas frontalement à la pensée formelle, elle est ce qui la produit. Il faut entendre toute l'ambiguité du terme de « conception » : comprendre une forme, c'est toujours se faire l'occasion de nouvelles formes (commentaires, remarques, questions, images). Et inversement, inventer (concevoir) une forme, c'est toujours partir d'autres formes. Concevoir, laisser l'intuition opérer, c'est se faire l'hôte de la prolifération des formes.

497 Henri Poincaré, *Science et Méthode*, Paris, Flammarion, 1947, p. 50-51.

498 Nous empruntons la comparaison à Arnaud Macé, « La naissance de la nature en Grèce ancienne », Haber et Macé (eds.), *Anciens et Modernes par-delà nature et société*, Besançon, Presses Universitaires de Franche-Comté, 2012, p. 47-84.

> Tout livre pousse sur d'autres livres, et peut-être que le génie n'est pas autre chose qu'un apport de bactéries particulières, une chimie individuelle délicate, au moyen de laquelle un esprit neuf absorbe, transforme, et finalement restitue sous une forme inédite non pas le monde brut, mais plutôt l'énorme matière littéraire qui préexiste à lui[499].

L'intuition est le naturel de la pensée au sens où, étymologiquement, naturel désigne ce qui naît et ce qui pousse[500]. Comprendre est moins « voir » que « faire naître une forme ». Nous sommes les hôtes de ces parasites (les formes) qui exigent notre travail pour se déployer. Nos esprits individuels forment un archipel de terroirs qui accueillent, font croître et circuler cette étrange *culture* consistant à faire naître des formes toujours plus hybrides. C'est bien une « culture » au sens où nous pouvons en favoriser certains produits, ou chercher à l'éliminer comme une mauvaise herbe, mais dont la croissance est spontanée.

> C'est avec la nécessité par laquelle un arbre porte ses fruits que poussent en nous nos pensées, nos valeurs, nos « oui » et nos « non, nos « si » et nos « que » - tous apparentés et reliés entre eux, et témoins d'une volonté, d'une santé, d'une terre, d'un soleil[501].

Cette « culture » effectue une circulation et un mélange des propriétés. Un mot de Proust, une couleur de Klein, viennnent inspirer un théorème mathématique, ou inversement. Ces influences formes un immense chaudron où s'effectue, en partie à notre insu, une cuisine des qualités.

Dans ce contexte, le mot de « culture » n'entre pas en opposition avec celui de « nature ». Au contraire, il y a culture au sens où il s'agit du naturel de la pensée, du principe de sa croissance. Cet intense trafic naît du vivant comme les plantes poussent de la terre. Dès lors, il paraît peu crédible d'autonomiser la pensée du vivant, aussi peu crédible que de coloniser Mars. Nos astronautes peuvent bien aller dans l'espace, même en sortie extra-véhiculaires, ils restent rattachés à la Terre et à sa « zone critique ». De la même manière, l'informatique semble éloigner du vivant le jeu de la culture, mais elle lui reste indissociable. Certaines machines, comme les navettes automatiques, peuvent faire preuve d'un peu d'« autonomie », elles restent entièrement dépendantes de nous, depuis leur conception jusqu'à leur recyclage, en passant par leur maintenance, leur mise à jour, et leur fourniture en énergie, en information, en passagers,

499 Julien Gracq, *Préférences,* Paris, José Corti, 1961, p. 82.

500 « […] *physis* dérive du verbe *phyesthai* 'croître' : la *physis* est ce qui croît et, avant de croître, naît. Le terme a été très justement rendu en latin par *natura*, dérivé de *nasci* 'naître'. », Jean-François Billeter, *Héraclite, le sujet,* Paris, Allia, 2022, Note II, p. 44.

501 Friedrich Nietzsche, *Pour une généalogie de la morale,* in *Œuvres,* Paris, Flammarion, 1996, p. 846-847.

etc. Surtout, plus le réseau s'étend, plus il dépend d'une multitude d'acteurs… humains. Aussi, issues du vivant, croissant de ce qui croît, nos idées peuvent s'incarner dans des formes artificielles semi « autonomes », elles ne s'émancipent pas pour autant du processus de la nature. Nous hébergeons ces formes comme on héberge un parasite et, comme eux, les formes ne survivent hors de leurs hôtes que le temps d'en infecter un nouveau.

Qu'il s'agisse d'énoncés, d'images, de théorèmes, d'idées politiques, mais aussi d'algorithmes, pour que les formes puissent conspirer à la naissance d'autres formes, il faut que des humains les *conçoivent.* Aucune nouvelle hybridation ne peut avoir lieu sans ce *rapport,* qui peut être amoureux et s'accompagner d'un certain plaisir. C'est également ce rapport qui permet leur conservation. La conservation d'une forme repose avant tout sur la conservation de son interprétation, la transmission de son feu qui donne le désir aux générations suivantes de poursuivre la conservation matérielle. Si on ne conçoit pas l'œuvre, si plus personne ne « voit » ce qu'elle « veut dire », elle tombe en décrépitude et s'efface[502]. Nos efforts de perpétuation n'offrent la « perpétuité » à aucune forme. Il faut toujours recommencer une transmission soigneuse et pleine d'efforts tout en sachant qu'au bout du compte, faute d'efforts ou faute d'humains, les traces finiront par s'effacer.

Rendre une machine pleinement autonome, la doter du même mouvement spontané que celui de la nature, c'est vouloir intégrer à son programme le principe qui fait naître et évoluer les choses (voir section 1.5.5. Machines autonomes, spontanées, voire naturelles). Nous pouvons *perpétuer* le vivant (voire l'aider à coloniser la matière inanimée), mais vouloir le *reproduire*, en abstraire et reproduire son principe de croissance, fait figure d'*hubris.*

2.6.5. Récapitulatif : une définition de l'intuition

L'intuition est l'« esprit réel[503] ». Comme la matière ou le temps, elle existe « sans pourquoi[504] ». L'intuition désigne la part de la raison qui est sans raison : les vérités premières, données brutes des sens, axiomes ou hypothèses[505]. C'est, selon les mots de Poincaré, le

502 Il n'est d'ailleurs pas nécessaire de savoir ce que l'œuvre signifiait dans son contexte. Il suffit que l'œuvre signifie, pour nous, en tant que trace des ancêtres par exemple.
503 C'est l'expression employée par Turing (« the real mind ») lors de l'exposition de l'analogie de l'oignon, dans l'article de 1950, voir section 1.5.5. Machine autonomes
504 Selon l'expression que Heidegger reprend de Silesius dans *Le principe de raison, op. cit.,* p. 102-109.
505 Selon la formulation de David Rabouin, voir supra.

« surgissement des idées[506] », l'origine des formes, tout ce qui vient à l'esprit et dont les formes constituées (images, sons, énoncés, algorithmes…) sont des traces plus ou moins pérennes – en fonction de leur inscription, ou non, sur des supports. Ce surgissement est spontané, mais il a besoin de nous, il s'effectue à la faveur de notre attention. La conscience que nous en avons forme le terreau sur lequel la « culture » peut se déployer. Notre interprétation des formes les fait résonner avec d'autres formes et leur permet de croître et de se perpétuer. Autrement dit nous les concevons : nous les comprenons, nous les engendrons, nous en ressentons du plaisir. Cette complicité que nous entretenons avec les formes constitue la vie de l'esprit ou le « naturel » de la culture au sens étymologique de « nature » : sa naissance et sa croissance spontanée.

C'est la même intuition qui préside au surgissement de nouvelles formes et à l'interprétation de formes déjà constituées. « Comprendre une idée » requiert la même intuition qu'« avoir une idée ». Dans les deux cas l'intuition est *partiellement involontaire*, « une pensée vient quand 'elle' veut et non pas quand 'je' veux[507] ; », bien qu'elle exige souvent, comme nous l'avons vu avec l'exemple donné par Poincaré, un effort conséquent ; *contingente*, elle peut avoir lieu ou non, sans raison particulière ; *éphémère*, les traces perdurent et permettent de renouveler l'idée, mais leur intuition est évanescente ; et *singulière,* elle est conditionnée par la personne, le lieu, le moment. Expérience singulière, l'intuition est aussi expérience de la singularité des formes et des situations. Elle nous permet d'appréhender en quoi elles ne sont comparables à rien d'autre – leur caractère exceptionnel.

L'intuition est ce qui permet de voir le sens d'une forme, d'une situation, mais aussi d'une opération. Elle est présente lorsque nous pensons à ce que nous faisons, et peut s'absenter aussi bien de nos actions manuelles que de nos opérations cognitives : nous pouvons penser à quelque chose *sans y penser* – sans penser au fait que nous y pensons. C'est le cas lorsque nous effectuons un algorithme sans prendre en compte le sens de l'opération (« pensée aveugle[508] ») ou lorsque nous suivons un enchaînement d'arguments (trame démonstrative) sans en effectuer l'interprétation pour nous (trame ascétique[509]). Ainsi, l'« esprit réel » ne se trouve pas en enlevant toutes les peaux de l'oignon de l'esprit mécanisable mais au sein de chacune des peaux

506 « les idées surgissaient en foule », Henri Poincaré, *Science et Méthode, op. cit.*, voir section 2.6.3. Concevoir une forme.
507 Friedrich Nietzsche, *Par-delà bien et mal, op. cit.,* p. 640.
508 C'est le terme inventé par Leibniz à ce sujet, voir supra.
509 Selon la distinction introduite par Foucault au sujet des *Méditations* cartésiennes, voir supra.

(des fonctions), selon que je l'effectue, ou non, en y pensant, en faisant, ou non, l'expérience du sens[510].

L'expérience du sens, autrement dit l'interprétation d'une forme ou d'une situation, se décline de nombreuses manières : il s'agit de percevoir ce à quoi elle renvoie en tant que signe (dénotation), ce qu'elle annonce (visée), ce qu'elle sert (finalité) et ce sur quoi elle se fonde (principes) – ce dernier étant le plus caractéristique de l'intuition puisqu'il s'agit de percevoir les principes ou fondements d'une forme ou d'une situation, autrement dit leur part originaire (sans raison). Cela requiert un mode de raisonnement particulier où la pensée revient sur ses pas vers des notions fondamentales comme le temps, la matière, la vie, mais aussi l'intuition et ses notions voisines (le sens, la compréhension…) : chacun fait l'expérience de ces concepts, en a une idée confuse, mais peine à les définir. L'intuition est ce qui permet de manipuler ces notions difficiles à concevoir sur lesquelles tout le reste du savoir repose. Malgré ses inconvénients majeurs – absence de fondement, incertitude, opacité, instabilité, variabilité inter et intra personnelle, difficultés à décrire… – nous ne pouvons pas faire l'impasse sur l'intuition, elle est le fond sans fondement à partir duquel tout le reste de la pensée naît.

Ainsi définie, l'intuition est un aspect de l'intelligence qui remet en cause la conjecture de Dartmouth, elle ne semble pas pouvoir être décrite d'une manière telle qu'une machine puisse la simuler. Ce n'est pas une opération qui s'explique en termes mécaniques puisque ce n'est pas une opération. Nous recevons une idée autant que nous la fabriquons, et souvent la réception passe par une suspension des idées déjà reçues où la pensée doit « reculer » au lieu d'avancer.

Une telle définition de l'intuition peut-elle clore le débat autour du projet d'intelligence artificielle ? Plusieurs questions demeurent. Tout d'abord, les réseaux de neurones peuvent-ils, depuis un parti pris empiriste selon lequel l'origine des formes, leur principe, seraient les « données brutes des sens », rendre compte, voire imiter certains aspects d'une telle notion de l'intuition ? D'autre part, le devenir des formes (la « culture ») pourrait-il s'émanciper de l'humain ? Pourrait-on cultiver les idées comme on cultive les virus en laboratoire, c'est-à-dire faire évoluer et muter les formes sans qu'un esprit n'ait à les concevoir ?

510 C'est une manière de reformuler l'argument de la chambre chinoise de Searle : l'opérateur de la chambre peut effectuer n'importe quelle opération et y prêter, ou non, attention.

2.7. L'intuition selon le *deep learning*

2.7.1. L'intuition : un réflexe visuel ?

Pour Yann LeCun, l'intuition est définie comme la capacité, basée sur l'accumulation d'expériences, à reconnaître instantanément une forme. Il donne l'exemple d'un joueur d'échecs si expérimenté qu'il peut jouer sans réfléchir.

> Je suis très mauvais aux échecs, mais je me suis trouvé un jour dans une situation passionnante : une quarantaine de personnes disputaient en même temps une partie contre un ancien champion de France. Celui-ci, bien sûr, ne passait pas plus de deux secondes avant de jouer un coup, et il a battu tout le monde à plate couture en quinze coups. Il s'agissait donc purement d'intuition. Il voyait l'échiquier et décidait en deux secondes. Il n'y avait plus aucun raisonnement de sa part[511].

Le journaliste qui mène l'interview invite Yann LeCun à préciser son propos : « Plus de raisonnement conscient. Mais un raisonnement quand même, dont il n'avait pas conscience. »

> C'est ça. Il avait intégré le jeu d'échecs dans son subconscient. Mais il aurait eu du mal à expliciter son raisonnement qui relevait presque de la perception visuelle pure. Si on demande à tout un chacun : « Comment reconnaissez-vous la lettre B ? », on peut évidemment la décrire, mais on a aucune idée du mécanisme de notre perception permettant d'arriver à ce résultat[512].

Il y a bien un raisonnement, mais celui-ci a été délégué « dans son subconscient » et peut difficilement être rendu explicite. On reconnaît le point de vue connexionniste pour qui les humains exécutent des tâches très complexes sans être capables de décrire le mécanisme à l'œuvre. En d'autres termes, la description subjective de ce qui se passe dans la tête du joueur ne présenterait pas d'intérêt. Les opérations intéressantes s'effectueraient en-dessous, « dans son subconscient ».

511 Yann LeCun in Stanislas Dehaene, Yann LeCun, Jacques Girardon, *La plus belle histoire de l'intelligence, Des origines aux neurones artificiels : vers une nouvelle étape de l'évolution, op. cit.,* p. 177.
512 *Ibid.*

Yann LeCun va plus loin et localise cette inscription subconsciente : c'est le système visuel qui a enregistré l'expérience du joueur d'échecs. « Son raisonnement » relève « presque de la perception visuelle pure ». Tout se passe comme si celle-ci fonctionnait comme un arc réflexe. A force de répéter une tâche, la raison *délègue au système visuel* la responsabilité de réagir à certaines situations bien identifiées. Grâce à l'accumulation d'expériences, le système visuel est capable d'associer une réponse à une situation sans qu'il n'y ait de réflexion consciente.

Dès lors, on comprend comment un programme conçu à l'origine pour de la vision artificielle et de la reconnaissance de forme en vient à être doté de la capacité à appliquer des stratégies. Un système perceptif peut « raisonner » au sens où il est possible de lui déléguer certains « raisonnements » préfabriqués par l'expérience – en l'occurence, un historique de milliers de parties de go jouées par des humains, auquel s'ajoutent des milliers de parties jouées par le programme. Cela revient à une théorie de l'intuition qui prend l'étymologie du mot au pied de la lettre : tout se passe au niveau de la vision.

D'après Yann LeCun, l'intuition fonctionne selon le processus suivant : lorsqu'une activité est inconnue, elle suscite un raisonnement conscient et « délibéré ». À ce stade, il s'agit « d'imaginer ce qui va se passer et de planifier une séquence d'actions qui va arriver à un résultat particulier. » Puis, une fois que l'action a été suffisamment répétée, « on compile ce comportement dans un sous-réseau de neurones », ce qui permet de la mettre à disposition sous forme d'automatisme. La tâche « devient intuitive » puisqu'il devient possible de « directement faire l'action sans avoir à réfléchir ». Il n'est plus nécessaire d'analyser la situation car une réponse toute faite est fournie instantanément. Autrement dit, l'intuition selon Yann LeCun consiste à fabriquer des automatismes à partir de la répétition de raisonnements conscients[513]. C'est cette fabrication d'automatismes que des systèmes comme AlphaGo reproduisent[514].

---

513 « quand on est pas habitué à faire une tâche particulière, en général on la fait de manière complètement raisonnée, délibérée, c'est-à-dire qu'on y pense. Par exemple, quand on joue aux échecs, qu'on est pas très bon comme moi, il faut qu'on explore les possibilités avant de jouer. Par contre on joue contre un grand maître qui lui peut jouer cinquante parties simultanées contre cinquante personnes; Il va toutes les gagner, lui il a pas besoin de réfléchir. Il voit l'échiquier, il joue immédiatement. C'est plutôt de la reconnaissance de formes et de l'intuition, que de l'exploration systématique.

Donc dans toutes les tâches qu'on peut faire, il y a cette possibilité de faire de l'exploration systématique, c'est à dire d'imaginer ce qui va se passer et de planifier une séquence d'actions qui va arriver à un résultat particulier. C'est complètement réflexif. Et puis, au fur et à mesure qu'on accomplit cette tâche on compile ce comportement dans un sous-réseau de neurones – si on peut parler en ces termes – qui est complètement réactif, c'est à dire qui peut directement faire l'action sans avoir à réfléchir. Et c'est très commun dans l'apprentissage humain. Et donc ensuite cette tâche devient intuitive. C'est à dire qu'on a plus besoin de faire la réflexion, de faire l'analyse pour voir si le résultat va être positif ou négatif, on a une intuition directe. » Entretien personnel avec Yann LeCun, 8 mars 2019.

514 « Et c'est un peu ce sur quoi se repose les systèmes de reconnaissance de forme et AlphaGo en particulier. Il joue des millions de parties contre lui-même. Qu'il finit par compiler dans un réseau de neurones qui est assez

L'idée générale est que l'accumulation d'expériences résulte dans le déclenchement de réponses automatiques à partir de la perception d'une situation. Cela peut passer par le système visuel aussi bien que par n'importe quel autre sens. Une réponse 'vient à l'esprit' de manière instantanée et directe. L'école connexionniste apporte ainsi une définition rigoureusement matérialiste de l'intuition : c'est un processus de mise au point d'automatismes se basant sur l'expérience ; tout s'effectue entièrement au niveau du corps puisque c'est le système perceptif qui reconnaît la situation et lui associe le bon automatisme.

2.7.3. Des yeux pour la « pensée aveugle »

L'approche connexionniste permet de rendre compte de certaines caractéristiques de l'intuition, à commencer par sa relation à la *perception*. Conformément au « goût philosophique » connexionniste, la pensée émerge depuis les « données brutes des sens[515] ». L'origine de la pensée se trouve dans la perception et, après un détour par l'élaboration consciente, elle y retourne. Une fois qu'un mécanisme est mis au point, les organes perceptifs reconnaissent les situations où il doit se déclencher et l'effectuent sans passer par les méandres de la délibération consciente. Cela permet d'expliquer l'*instantanéité* de l'intuition : elle fonctionne à la manière d'un « arc réflexe », comme une réaction automatique allant directement du stimuli à la réponse ; son *opacité* : le produit de l'intuition (image, énoncé, geste…) peut être clair, mais je ne sais pas d'où il vient, ni comment il a été produit ; ainsi que son aspect *involontaire* : l'association entre la situation et le mécanisme déclenché se fait malgré moi et je ne peux intervenir qu'à la fin du processus, en décidant de ne pas donner suite à ce que l'intuition m'invite à effectuer.

Enfin, il y a une ébauche d'explication du mode « oraculaire » de l'intuition, c'est-à-dire la capacité à me fournir une réponse avant d'avoir une idée précise de la situation ou du problème. Au moment du match entre AlphaGo et Lee Sedol, le go a été présenté comme étant d'un niveau de complexité tel que seule l'intuition permettrait d'y jouer. Si le jeu contient plus de combinaisons qu'il n'y a d'atomes dans l'univers, il est impossible d'avoir une idée précise de la situation et de jouer en s'appuyant sur une méthode procédurale. Il est nécessaire de

gros, qui est un réseau convolutif, qui regarde l'échiquier et qui produit un coup. Bon il y a un petit peu d'exploration arborescente mais à la fin il n'y en pas besoin de beaucoup ; et qui finalement compile tout ce savoir dans un réseau intuitif, qui a l'intuition de jouer le bon coup. » *Ibid.*

515 Nous reprenons ici la formulation de David Rabouin, voir supra.

recourir à l'intuition. Seule l'expérience, et pour AlphaGo la répétition de millions de parties en *self-play,* permet d'associer une réponse aux situations rencontrées.

Parler d'intuition au sujet d'AlphaGo et des réseaux de neurones en général permet donc d'avancer plusieurs hypothèses intéressantes, notamment au sujet de l'origine empirique des idées, de la continuité entre perception et pensée, et de l'efficacité permise par l'accumulation d'associations tissées au gré de l'expérience. Mais, dans la mesure où les réseaux de neurones sont un assemblage d'algorithmes, ils restent exposés à la critique de Searle. Chacune de leurs briques constitutives sont des procédures simples d'où la pensée peut être évacuée[516]. On ne voit pas « où » ni « comment » pourrait avoir lieu une conscience de ce qui est effectué, c'est-à-dire une compréhension des opérations qu'ils effectuent. L'apprentissage profond est capable de proposer un modèle pertinent de la vision des objets réels, mais pas de « voir » ce que signifie une proposition, c'est-à-dire d'y percevoir un sens. AlphaGo s'inspire du système visuel mais ne « voit » rien. Tout se passe comme si la catachrèse « la pensée a des yeux » était prise en un sens littéral, et, dans un souci de matérialisme, renversée au point d'affirmer que « l'intuition est de la pensée *effectuée par les yeux »,* mais dans la continuité de la pensée aveugle : « l'intuition est de la pensée effectuée par les yeux *sans que ceux-ci ne la voient* en tant que pensée ».

L'intuition est réduite à une action répétée qui, à force de répétition, serait vidée de sa conscience[517]. L'intuition selon le *deep learning* est de la pensée aveugle, encore plus aveugle que la manipulation de symboles mathématiques puisqu'elle s'effectue à l'insu du penseur. Elle consiste en la matérialisation de raccourcis de pensée dans des « sous-réseaux de neurones ». Autrement dit c'est une boîte à outils, un ensemble de « trucs » ou d'artifices pour penser plus vite – une prothèse constituée à l'intérieur du cerveau. Elle est aussi mécanique que les raisonnements explicites basés sur des règles formelles, et encore plus automatique puisqu'elle ne laisse aucune place à la délibération. Nous voilà loin de la « lumière naturelle » historiquement associée à l'intuition. L'idée que le système nerveux incorpore des automatismes cognitifs se trouve déjà chez Descartes, sans que cela l'empêche de développer une notion distincte d'intuition[518].

---

516 Voir supra 1.5.7. Reconnaître l'intelligence.

517 « Qu'arrive-t-il quand une de nos actions cesse d'être spontanée pour devenir automatique ? La conscience s'en retire. Dans l'apprentissage d'un exercice, par exemple, nous commençons par être conscients de chacun des mouvements que nous exécutons, parce qu'il vient de nous, parce qu'il résulte d'une décision et implique un choix, puis, à mesure que ces mouvements s'enchaînent davantage entre eux et se déterminent plus mécaniquement les uns des autres, nous dispensant ainsi de nous décider et de choisir, la conscience que nous en avons diminue et disparaît. » Henri Bergson, *L'énergie spirituelle,* Paris, Presses universitaires de France, 2017.

518 David Bates, « Cartesian Robotics », *Representations,* 124, 2013, p. 47-48.

À cela s'ajoute que le fait d'arrimer l'intuition à l'expérience échoue à rendre compte pleinement de la dimension « oraculaire » évoquée par Turing en 1938[519]. Si l'intuition est constituée par l'accumulation d'expériences, on ne voit pas comment elle peut s'appliquer à un problème nouveau. L'intuition selon le *deep learning* fournit une explication intéressante de ce qui s'effectue lorsque, à force d'expérience, le savoir s'est constitué en automatismes, mais elle se réduit à des *réflexes.* Elle échoue à rendre compte de l'invention, des conjectures et des hypothèses – de ces moments qui ne peuvent (sauf à adopter une théorie de la réminiscence platonicienne) être interprétés comme la reconnaissance de formes *déjà vues*. L'explication de Yann LeCun restitue avec élégance ce qui se passe quand on mobilise un savoir passé – dont on a oublié la genèse – mais laisse de côté ces connaissances *à venir* – dont *on ne sait pas encore* comment on les connaît. Il ne rend pas compte d'une éventuelle capacité de la pensée à *concevoir :* à inventer et à comprendre. Ce sont pourtant des propositions de ce type qui motivent les conjectures et donc le geste fondateur du projet d'intelligence artificielle qu'est l'hypothèse formulée à Dartmouth. En d'autres termes, avoir l'intuition d'un modèle réflexe de l'intuition est paradoxal, puisqu'un tel modèle ne rend pas compte pleinement de l'existence de l'intuition.

2.7.4. L'intuition comme superstition ou illusion

Si leurs hypothèses de travail sont différentes, l'ambition des chercheurs en *deep learning* reste fidèle à celle qui a présidé à la genèse du projet d'intelligence artificielle – à savoir la promesse d'une description exhaustive des mécanismes de l'intelligence – qui implique de faire un sort à l'intuition. Il s'agit d'apporter une théorie de l'intuition équivalent à une description « si précise qu'une machine peut la simuler », autrement dit d'en écrire l'algorithme. Mais postuler qu'on peut écrire l'algorithme de l'intuition revient à réduire celle-ci à un sous-ensemble de la pensée mécanisable, à une « peau » de l'oignon. La distinction entre pensée formelle et intuition, ou « pensée aveugle » et pensée « voyante », perd toute consistance puisque l'ensemble de la pensée peut être appréhendée par la pensée formelle. En perdant cette spécificité, l'intuition devient difficile à définir : qu'est-ce qui la distingue du reste de la pensée ? S'il ne s'agit que d'automatismes de pensée acquis par expérience, en quoi se distingue-t-elle des réflexes ? Alors que le premier projet d'IA ignorait purement et simplement l'intuition, le connexionnisme a

519 Alan Turing, « Systems of Logic Based on Ordinals », in Jack Copeland (ed.), *op. cit.*

pris le parti de l'intégrer, mais de si bien l'intégrer qu'on ne voit plus en quoi elle se distingue de la pensée formelle, ni ce qui la définit. En d'autres termes, le connexionnisme reprend, de façon détournée, l'objectif d'élimination de l'intuition. Le concept n'aurait eu droit de cité qu'à cause d'un mysticisme appliqué à la pensée, mysticisme qu'il conviendrait de mettre au placard des superstitions obsolètes.

On observe un étrange renversement. À l'époque de Leibniz la pensée formelle était vue comme un moment particulier de la pensée où celle-ci joue momentanément à être aveugle, à se priver d'une partie de ses facultés qui ressemble à la vue. À notre époque, c'est au contraire l'intuition qui est vue comme un aveuglement, comme une superstition. La pensée s'aveugle quand elle mobilise le concept d'intuition. Elle s'invente des moments inexplicables de façon à jeter un voile pudique sur ses propres mécanismes, trop prosaïques pour être à la hauteur de la créativité qu'elle revendique. Alors qu'au dix-septième siècle la pensée est « aveugle » quand elle *joue à la machine*, au vingt-et-unième siècle, la pensée est « aveugle » – elle s'aveugle – quand elle se raconte qu'elle *n'est pas une machine*. Au dix-septième, « l'œil de la raison » est l'état permanent de la pensée dont on se prive momentanément pour faire des calculs dans le but d'arriver à une invention. Au vingt-et-unième siècle, « l'œil de la raison » est une illusion que l'on convoque pour s'attribuer une invention.

Si toute la pensée peut être décrite en langage formel, voire si toute la pensée est considérée comme du calcul sur un langage symbolique, alors l'ensemble de la pensée est formelle sans le savoir. Toute la pensée est « aveugle », et elle se leurre quand elle pense y « voir » du sens ou « choisir » librement une direction de réflexion. Les manifestations de l'intuition ne sont que des calculs qui s'ignorent. La pensée ne sait pas qu'elle est aveugle, elle s'aveugle au sujet de son aveuglement et ne peut « voir » qu'une seule chose : qu'elle ne voit rien.

2.7.5. AlphaGo invente de nouveaux coups

Lors du deuxième match, un coup joué par AlphaGo provoque la stupéfaction du public. Selon l'un des commentateurs, le coup « est mauvais. Nous ne pouvons pas dire pourquoi, il est tout simplement mauvais[520] ». En coulisses, l'équipe de programmeurs se demande s'il n'y a pas eu une erreur. Seul Lee Sedol semble positivement surpris. « Ce coup était vraiment magnifique

520 « It's bad. We don't know why, simply it's bad », déclaration de Fan Hui filmé par Greg Kohs, *op. cit.*

et créatif[521] » dira-t-il après le match – et une nouvelle victoire d'AlphaGo. Cela est d'autant plus remarquable que Lee Sedol est célèbre pour son style audacieux et revendique le fait de jouer d'une façon radicalement nouvelle.

D'après les statistiques calculées par un sous-programme d'AlphaGo, un humain sur dix mille aurait joué le coup 37 du deuxième match contre Lee Sedol. Tout comme les machines de Turing le « prennent par surprise très fréquemment », AlphaGo réussit à surprendre jusqu'à ses propres programmeurs. Pour autant, est-il pertinent de considérer qu'AlphaGo a *conçu* quelque chose de nouveau [originated] et ainsi répondu à l'objection de Lovelace ? Si les ingénieurs sont surpris par le déroulement du match, c'est avant tout parce qu'ils connaissent mal le jeu de go. Ils sont incapables d'évaluer par eux-mêmes les coups proposés. Comme le rapporte Turing au sujet de ses propres surprises, ils ne font « pas assez de calculs » pour pouvoir prévoir le cours des choses. Le phénomène se reproduit pendant une partie du cinquième match : alors que toute l'équipe est persuadée que le programme dysfonctionne, des coups atypiques lui permettent finalement de l'emporter.

D'après un des commentateurs, AlphaGo « est allé au-delà de son guide humain et a créé quelque chose de nouveau et de différent[522] ». Pourtant, selon un des ingénieurs de l'équipe, « AlphaGo est vraiment un programme simple[523] ». Dans la mesure où il s'agit d'un programme, c'est-à-dire d'une manipulation automatisée de symboles selon des règles de syntaxe strictes, en quoi la 'création' d'AlphaGo est-elle différente des « lutins » et des « fées » de Lady Lovelace, c'est-à-dire des surprises opérées par la manipulation « aveugle » de formes symboliques ? À l'instar de l'*ars inveniendi* qu'est l'algèbre, la manipulation autonome de formes opérée par les algorithmes permet de mettre au jour des connaissances difficiles d'accès, quand l'objet à manipuler dépasse les capacités de notre esprit – repoussant ainsi l'horizon de la pensée. On se souvient de l'éloge de la « pensée aveugle » par Leibniz, celle-ci permettant d'appréhender les objets inimaginables comme un polygone à mille côtés. Aussi arrive-t-on à un point où la distinction entre intuition et pensée aveugle se trouble. Si j'utilise un langage formel pour manipuler (ou faire manipuler par une machine) des objets dont la complexité dépasse mon entendement, produisant ainsi des résultats « nouveaux » ou tout du moins « surprenants », il s'agirait pour l'école connexionniste, d'intuition, alors que pour Leibniz, il s'agirait de son contraire (pensée aveugle). De ce point de vue, quelle pertinence y-a-t-il à attribuer *à AlphaGo* la création de « quelque chose de nouveau et différent » ? Ne s'agit-il pas

521 « This move was really creative and beautiful », déclaration de Lee Sedol, *Ibid.*
522 « it went beyond its human guide and created something new and different », *ibid.*
523 « AlphaGo is really a simple program », *ibid.*

seulement d'un instrument de découverte (*ars inveniendi*) au service des humains qui l'utilisent ?

À la suite de ses multiples défaites, Lee Sedol se remet en question dans des termes qui rappellent la réponse de Turing à l'objection de Lady Lovelace : pour quelles raisons a-t-il pu penser qu'il est un joueur créatif ? Plus généralement, qu'est-ce qui nous permet de considérer que nous sommes capables de création ? Notre cerveau ne fonctionne-t-il que comme une machine déterminée et nous nous leurrons, comme le suggérait Turing en 1950, lorsque nous pensons être « créatifs » ou originaux ? Ou bien existe-t-il une part de notre pensée dont on puisse dire qu'elle a la capacité d'engendrer des formes nouvelles ? Si cette « part » existe, sommes-nous capables de la décrire, et de la décrire si précisément qu'une machine serait capable de la simuler ?

TROISIEME PARTIE

MECANISME ET CREATION

3.1. Avoir du jeu : la paidia

3.1.1. La créativité à l'insu du sujet

Dans un autre passage célèbre de *Science et méthode,* Poincaré raconte comment lui apparaît la solution d'un problème mathématique. Après avoir longuement cherché, il s'interrompt pour participer à « une course géologique » et oublie ses travaux, tout absorbé par les « péripéties du voyage ». C'est au moment de monter dans un omnibus que « l'idée [lui] vînt », « sans que rien dans [ses] pensées mathématiques parût [l'y] avoir préparé[524] ». Tout se passe comme si une partie de lui avait travaillé *à son insu* pour ne se manifester qu'une fois la solution disponible – c'est le moment du « Eurêka ».

De tels exemples sont nombreux dans l'histoire de la pensée. La solution vient d'un coup, à un moment où l'esprit ne travaille pas consciemment au problème, comme si elle arrivait d'ailleurs. C'est allongé dans l'herbe, après une longue course à pied, que Turing conçut l'idée directrice de l'article de 1936[525]. George Boole eut l'intuition d'appliquer l'algèbre à la logique alors qu'il marchait dans un champ – un épisode que son biographe compare à l'illumination de Saint-Paul en chemin vers Damas[526]. C'est pendant un trajet en bateau que Leibniz élabore l'idée de la caractéristique universelle[527]. Ou encore, c'est en composant un poème d'amour

524 « A ce moment, je quittai Caen, que j'habitais alors, pour prendre part à une course géologique entreprise par l'École des Mines. Les péripéties du voyage me firent vite oublier mes travaux mathématiques ; arrivés à Coutances, nous montâmes dans un omnibus pour je ne sais quelle promenade ; au moment où je mettais le pied sur le marchepied, l'idée me vint, sans que rien dans mes pensées mathématiques parût m'y avoir préparé, que les transformations dont j'avais fait usage pour définir les fonctions fuchsiennes étaient identiques à celles de la Géométrie non euclidienne. » Henri Poincaré, *Science et Méthode*, Paris, Flammarion, 1947, p. 51.

525 « It had become his habit to run long distances in the afternoons, along the river and elsewhere, even as far as Ely. It was at Grantchester, so he said later, lying in the meadow, that he saw how to answer Hilbert's third question. It must have been in the early summer of 1935. 'By a mechanical process', Newman had said. So Alan Turing dreamed of machines. » Andrew Hodges, *Alan Turing : The Enigma*, Princeton, Princeton University Press, 2014, page 123.

526 « He also liked to speak of the inspiration that suddenly came to him during his stay at the Methodist school. While walking across a field, the thought flashed across his mind that it should be possible to express logical relationships in algebraic form. This experience, which a biographer compares to that of Paul on the road to Damascus, was to bear fruit only many years later. » Martin Davis, *The Universal Computer, The Road from Leibniz to Turing*, New-York, W. W. Norton & Company, 2000, page 24.

527 « En voulant aller d'Angleterre en Hollande j'ay esté retenu quelque temps dans la Tamise par les vents contraires. En ce temps la ne sçachant que faire et n'ayant personne dans le vaisseau que des mariniers, je méditois sur les choses-là et surtout je songeois à mon vieux dessein d'une langue ou écriture rationnelle, dont le moindre effet serait l'universalité et la communication de différentes nations. » Lettre à Gallois 1677, *in* G.

que Lulle est frappé par l'idée de l'Ars[528]. Dans tous ces exemples, qui sont, pour reprendre les termes de Turing en 1938, autant de manifestations de « jugements spontanés », il y a un effort conscient du sujet mais la solution ou l'idée n'apparaît qu'après un moment de distraction qui semble permettre que s'effectue un processus autonome, processus qui ne devient conscient qu'après-coup, au moment de livrer le résultat[529]. Tout se passe comme s'il fallait *faire autre chose* pour que le processus puisse aboutir. Ce dernier n'aurait lieu qu'à condition que le sujet détourne le regard et se garde d'intervenir, de la même façon que l'exercice de l'algèbre requiert une suspension de l'attention pour éviter de perturber la transformation des formes mathématiques. Est-ce qu'à l'instar de la « pensée aveugle » selon Leibniz, la créativité requiert de se bander momentanément les yeux pour déléguer la manipulation des formes à une sorte de machine intérieure ?

Dans le passage déjà commenté où il décrit comment un bouillonnement d'idées provoque des accrochages et se stabilise en une combinaison finalisée qui l'amène à la découverte des fonctions fuschiennes[530], Poincaré semble avoir assisté au processus habituellement inconscient qui précède l'irruption de la découverte. C'est un processus non dirigé où par essai-erreur un maximum de combinaisons possibles sont testées. Poincaré s'y est appliqué consciemment (« j'essayais un grand nombre de combinaisons ») sans aboutir. Peut-être faut-il, pour que la solution émerge, qu'il *se retire,* qu'il suspende sa volonté, de façon à laisser la place à une plus grande variété de combinaison de formes – comme si l'activité consciente du sujet entravait l'élaboration de la découverte. S'il assiste au processus, Poincaré

W. Leibniz, *Leibnizen Mathematische Schriften, éd.* C. I. Gerhardt, Berlin, Verlag von Asher & Comp., 1849, p. 180-181.

528 Pour Lulle il s'agit plutôt d'une inspiration subite qui semble sans rapport avec la tâche qu'il cherchait à exercer consciemment. Le Christ lui apparaît plusieurs jours d'affilée, à chaque fois qu'il cherche à achever un poème d'amour. Lui vient alors l'idée qui guidera l'invention de l'Ars : « composer un livre qui serait le plus efficace du monde contre les erreurs des infidèles. » voir Josep Rubio, *Raymond Lulle, le langage et la raison : une introduction à la genèse de l'Ars,* Paris, Vrin, 2007, p. 15-16. Pour Lulle, ce que nous appellerions aujourd'hui la fécondité de la pensée formelle (via l'art combinatoire) et l'inspiration subite sont une seule et même chose : émanations de la pensée de Dieu. Si bien que « Dieu est l'auteur, en dernier lieu, de l'*Ars* lullienne. » *Ibid, p.* 17.

529 L'idée que la créativité s'effectue à l'insu de la conscience a été largement débattue au dix-neuvième siècle, suite à l'invention de la notion d'inconscient, ainsi qu'aux travaux de psychologues qui décrivent la pensée comme étant principalement une activité nerveuse involontaire. David Bates, « Insight in the Age of Automation », *in* Joyce Chaplin, Darrin McMahon (dir.), *Genealogies of Genius,* Londres, Palgrave Macmillan, 2016.

530 Pour rappel : « tous les jours, je m'asseyais à ma table de travail, j'y passais une heure ou deux, j'essayais un grand nombre de combinaisons et je n'arrivais à aucun résultat. Un soir, je pris du café noir, contrairement à mon habitude, je ne pus m'endormir : les idées surgissaient en foule; je les sentais comme se heurter, jusqu'à ce que deux d'entre elles s'accrochassent, pour ainsi dire, pour former une combinaison stable. Le matin, j'avais établi l'existence d'une classe de fonctions fuchsiennes, celles qui dérivent de la série hypergéométrique ; je n'eus plus qu'à rédiger les résultats, ce qui ne me prit que quelques heures. » Henri Poincaré, *op. cit.,* p. 50-51.

se garde bien d'intervenir : ce sont les idées qui « surgissent en foule », se heurtent et s'accrochent, tandis qu'il reste passif.

Les idées s'agencent d'elles-mêmes comme si elles étaient autonomes. Mais Poincaré n'utilise pas de machine, ni de feuille, ni de crayon. Les idées dépendent de son esprit pour pouvoir se manifester. Il y a là une sorte d'hospitalité : une part de son esprit lui devient étranger et accueille des propositions de formes. Il se laisse embarquer, occuper, posséder, par une idée que Poincaré comme sujet conscient ne se permettrait pas ou n'aurait pas les moyens d'avoir. Quelque chose comme « un autre en lui[531] » pense pour lui et permet de déjouer la censure qui l'empêche de voir certains liens pertinents. En d'autres termes, « l'autre en lui » aboutit à étendre son horizon de pensée. Le fonctionnement ressemble à celui d'une prothèse, mais sans support matériel extérieur. Cet « autre en lui » est-il une machine ?

Tout se passe comme si, pour avoir une idée, il fallait renoncer à contrôler ses pensées et se 'laisser embarquer' par un jeu spontané des formes, un *ars inveniendi* involontaire. L'algèbre, telle qu'elle est décrite par Leibniz, ne serait qu'un exemple d'un processus plus général, consistant à laisser les formes se combiner spontanément, qui requiert parfois de se bander momentanément les yeux (passer par une « pensée aveugle ») pour que viennent des pensées intéressantes. Dès lors, la distinction entre « pensée aveugle » et intuition (au sens de surgissement des idées) se brouille. L'émergence de nouveauté semble provenir d'un jeu autonome des formes plutôt que d'un effort du sujet. Le sujet ne serait pas « créatif » mais une entrave à l'émergence de nouveauté. Il faudrait une *suspension* du sujet pour laisser libre cours à un processus passant par une machine extérieure (*ars inveniendi*) ou intérieure (« autre en soi »). Lors de l'insomnie de Poincaré, les formes se combinent *d'elles-même*, par essai-erreur, et arrivent à de nouveaux agencements. Nous retrouvons la question que pose Turing en objection à Lady Lovelace : en quoi l'humain est-il « créateur » des formes ? L'invention de nouvelles formes pourrait-elle avoir lieu *sans nous* ? La créativité pourrait-elle *se passer complètement* de sujet et être effectuée par une machine ?

531 Nous empruntons cette notion à Pierre Cassou-Noguès qui explore, dans son étude sur Gödel, le thème de l'inspiration comme une altérité du soi, et le relie à la volonté de compléter l'arithmétique. Pour compléter l'arithmétique, il faudrait arriver à « penser sans son cerveau », accueillir au sein de sa pensée des pensées « dont le cerveau n'est pas capable ». Voir Pierre Cassou-Noguès, *Les démons de Gödel, op. cit.*, p. 158-159.

3.1.2. S'émanciper de l'expérience humaine, d'AlphaGo à AlphaZero

Suite au succès d'AlphaGo, les ingénieurs de DeepMind ont créé un nouveau programme nommé AlphaGo Zero dont l'apprentissage ne passe pas par l'assimilation d'un historique de dizaines de milliers de parties. AlphaGo Zero s'entraîne exclusivement via un apprentissage par renforcement. Le programme se paramètre en jouant plusieurs millions de parties virtuelles et retient les stratégies qui amènent à la victoire. Il s'agit d'explorer le plus grand nombre de combinaisons possibles, par essai-erreur, sans l'apport d'aucune expérience humaine. Après un entraînement suffisant, AlphaZéro l'emporte sur AlphaGo[532].

Pour Demis Hassabis, le nouveau programme est meilleur parce qu'il « n'est pas restreint par les limites de la connaissance humaine[533] ». L'apprentissage initial d'AlphaGo à partir d'un historique de parties l'aurait cantonné à une imitation des humains et empêché d'explorer toutes les combinaisons possibles. Jouer à la manière des humains, c'est aussi reproduire leurs préjugés, alors que la machine est neutre, elle est capable de traiter toutes les combinaisons de la même façon. Cette absence de préférences faciliterait la découverte de stratégies plus efficaces – à l'instar des « suprises » engendrées par AlphaGo.

La supériorité d'AlphaGo Zéro au jeu de go, puis aux échecs et au shogi auxquels il est successivement appliqué, semble démontrer que le choix d'une bonne stratégie n'est qu'une affaire de combinatoire, et donc de puissance de calcul. La machine bénéficierait d'une double supériorité sur l'humain, d'une part, lorsqu'elle est dotée de beaucoup de puissance de calcul, en étant capable d'effectuer *plus* de combinaisons, et d'autre part en testant toutes les combinaisons sans qu'aucun préjugé ne vienne la gêner. Est-ce qu'un tel mécanisme reproduit bien le processus de créativité tel qu'il est décrit par Poincaré ? Le programme AlphaGo Zero peut-il être qualifié de machine créative, ou d'*imagination artificielle,* capable de reproduire ce qu'effectue l'imagination humaine ?

---

532 David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Hui Fan, Laurent Sifre, George van den Driessche, Thore Graepel et Demis Hassabis, « Mastering the game of Go without human knowledge », *Nature*, vol. 550, no 7676, 19 octobre 2017, p. 354–359

533 « DeepMind co-founder and CEO Demis Hassabis said the programme was so powerful because it was "no longer constrained by the limits of human knowledge". » Sarah Knapton, « AlphaGo Zero: Google DeepMind supercomputer learns 3,000 years of human knowledge in 40 days », *The Telegraph*, 18 octobre 2017, https://www.telegraph.co.uk/science/2017/10/18/alphago-zero-google-deepmind-supercomputer-learns-3000-years/ consulté le 14 septembre 2019.

3.1.3. À quoi joue AlphaGo Zero ? *Ludus* et *paidia*

Nous avons déjà souligné comment l'algèbre et plus généralement l'opération d'une machine peuvent être utilisés comme des instruments de découverte. Le fait que nous puissions déléguer une partie du processus créatif vers de tels outils mécaniques laisse penser que la création est, au moins en partie, déjà un processus mécanique intériorisé. Le récit de Poincaré amène à penser que processus créatif et *ars inveniendi* ont en commun de s'effectuer dans une relative autonomie par rapport à un opérateur humain. Un contenu s'y forme et s'y transforme selon un système de règles propres. A cela s'ajoute qu'il vaut mieux détourner le regard du processus pour ne pas l'entraver. Aussi, alors qu'il est fréquent que la créativité soit brandie comme l'activité qui distingue l'humain de ses machines, les points communs entre le processus créatif et l'opération d'une machine sont nombreux.

Toutefois, la « créativité » d'AlphaGo Zéro est dépourvue de réflexivité. Son *absence de jugement* est différente de la *suspension du jugement* qui laisse libre cours au jeu spontané des formes, cette dernière étant l'occasion d'un retour sur les principes, d'un déverrouillage des règles du jeu permettant d'en démultiplier les possibilités. Grâce à son absence de préjugés, le programme explore sans contraintes la totalité des combinaisons du jeu, il fait preuve d'une exhaustivité dont l'humain est souvent incapable. Mais il ne peut *ouvrir* la gamme de possibilités qu'il explore, autrement dit, appliquer l'exploration au jeu lui-même. Le programme ne fait pas de proposition *au sujet* de l'opération qu'il effectue. Il ne pourrait, pour reprendre les termes de Lovelace, « être à l'origine » d'un changement des règles du jeu, ni s'adapter de lui-même à un changement de définition du jeu. Si la taille du *goban* venait à changer, le programme devrait recommencer son entraînement pour retrouver une bonne calibration et pouvoir jouer à nouveau.

Pour Roger Caillois, le jeu en tant qu'activité disciplinée par des règles, « combinaison et calcul », qu'il appelle *ludus*, est produit par une notion plus générale du jeu, la *paidia* ou « fantaisie sans règle[534] ». Celle-ci est à l'œuvre chez l'enfant qui a « le goût d'inventer des règles et de s'y plier obstinément, quoi qu'il en coûte ». Il lance « toutes sortes de paris, il marche à cloche-pied, à reculons, en fermant les yeux, joue à qui, le plus longtemps, regardera le soleil, supportera une douleur ou demeurera dans une position pénible[535]. »

534 Roger Caillois, *Les jeux et les hommes, Le masque et le vertige*, Paris, Gallimard, 1967, p. 75-91.
535 *Ibid*, p. 78-79.

Pour chaque jeu les règles sont explicites, mais quelles règles régissent le passage d'un jeu à un autre ? Une fois qu'une règle est définie, un programme peut l'appliquer aussi « obstinément » que le joueur humain, mais il manque au programme ce sens de la *paidia* qui permet à deux enfants de transformer les règles du jeu pendant qu'ils jouent. AlphaZero peut essayer toutes les combinaisons au sein d'un jeu donné mais il ne peut effectuer de lui-même la transition vers un nouveau jeu. Si un programme peut improviser *selon* des règles, il ne peut improviser de nouvelles règles.

La *paidia* selon Roger Caillois est un cas où la pensée ne peut être formalisée. Une combinatoire fermée suffit pour s'attaquer au *ludus* comme le go, qui a des règles bien définies. Mais pour s'appliquer à la « fantaisie sans règles » ou *paidia*, le champ de la combinatoire doit s'étendre aux règles du jeu. Il faut pouvoir tester des combinaisons définies au sein de l'ensemble de règles mais aussi jouer sur ces règles. Or, le modèle de « créativité » que propose un programme comme AlphaGo Zero échoue à restituer cette dimension.

3.1.4. L'invention scientifique selon Bachelard

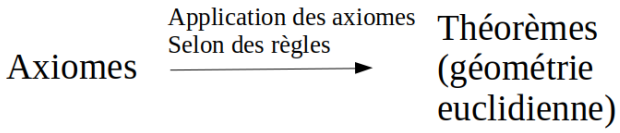
L'invention scientifique requiert sa propre *paidia*, un jeu sur les règles dont Bachelard nous donne une illustration avec le cas de l'invention de la géométrie non euclidienne par Lobatchewsky. Au lieu de se contenter de dérouler des théorèmes à partir des axiomes et de participer à un enrichissement « monotone » de la géométrie, Lobatchewsky inclut les axiomes dans sa réflexion et choisit de les modifier. Il ne transforme pas selon des règles, il transforme les règles elles-mêmes et aboutit à un autre système cohérent.

> la somme des angles d'un triangle est égale à deux droits. Vous lui répondez tranquillement : « ça dépend. » En effet, cela dépend du choix des axiomes. D'un sourire, vous déconcertez cette raison tout élémentaire qui s'accorde le droit de propriété absolue sur ses éléments. Vous assouplissez cette raison dogmatique en lui faisant jouer de l'axiomatique[536].

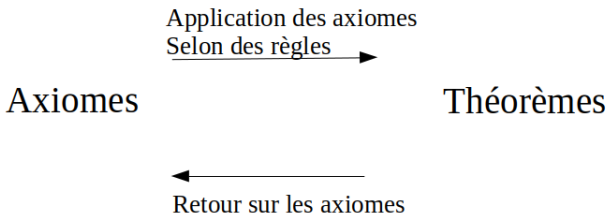
Alors qu'une machine limite son « jeu » aux combinaisons de formes encadrées par un ensemble de règles, la raison humaine étend le « jeu » aux règles elles-mêmes. Elle ne pense

536 Gaston Bachelard, *L'engagement rationaliste,* Paris, PUF, 1972, p. 9-10.

pas pour autant sans principes : la géométrie se fonde toujours sur un système d'axiomes, mais ceux-ci ont été modifiés. Il y a un premier mouvement qui consiste à explorer l'espace des possibilités offert par un jeu d'axiome, exploration qui peut être confiée à une machine, comme le montre l'illustration suivante :



Il y a un deuxième mouvement qui consiste à revenir sur les axiomes. Peut-on suspendre tel ou tel axiome ? Par quel autre axiome faudrait-il le remplacer ? Quel en serait l'effet ?

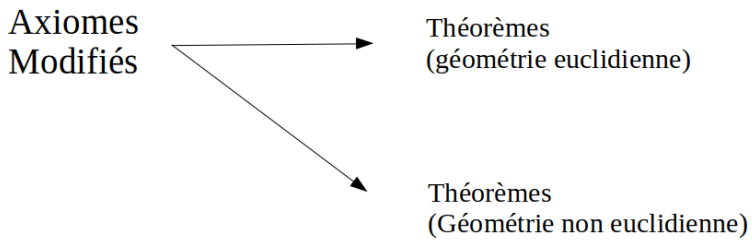


Une forme de *paidia* a lieu quand Lobatchewsky suspend l'axiomatique euclidienne et adopte un nouvel axiome. Pour que ce jeu sur les axiomes soit formalisable, et donc mécanisable, il faudrait que l'espace des axiomes soit défini, que soit cartographiées les axiomes pouvant être modifiés et les modifications possibles. Mais à ce moment précis, il n'existe pas de règle selon laquelle modifier les règles : c'est un « délire » au sens étymologique du mot – une « sortie du sillon[537] » – sans être pour autant une folie[538] : cette sortie du sillon de la géométrie euclidienne constitue aussitôt un nouveau sillon, aussi cohérent (et qu'une machine pourra continuer à creuser). Bachelard utilise un vocabulaire révolutionnaire, il évoque une « fonction de

537 Félix Gaffiot, *op. cit.*, p. 491.

538 Pour David Bates, l'erreur, ou plus précisément l'errance, est « le signe de la capacité de l'esprit à découvrir quelque chose de nouveau », David Bates, « Automatisation et erreur », in Bernard Stiegler (ed.), *La verité du numérique : Recherche et enseignement supérieur à l'ère des technologies numériques,* Paris, FYP Éditions, 2018, 29-40. Voir aussi David Bates, *Enlightenment Aberrations : Error and Revolution in France,* Cornell University Press, 2002.

turbulence » de la raison qui rappelle le « tumulte et exubérance[539] » de la *paidia* selon Caillois (opposée au « calcul et combinaison » du *ludus*[540]).



Une fois les axiomes renouvelés, la géométrie euclidienne – qui était confondue avec la géométrie en général – devient un sous-ensemble d'une géométrie plus vaste et doit cohabiter avec de nouvelles géométries. L'invention de Lobatchewsky ouvre à un pluralisme géométrique, à charge ensuite pour la raison de trouver quels principes élémentaires rendent compte de l'ensemble des géométries[541].

Une « raison prudente » serait restée dans le cadre défini par les axiomes d'Euclide et se serait cantonnée à l'élaboration de nouveaux théorèmes à partir de ce socle et des théorèmes déjà prouvés – à l'instar d'une machine qui, par construction, ne peut que dérouler les conséquences du système de règles. Lobatchevski se permet un moment de réflexivité. Il prend du recul par rapport aux axiomes, il se place en amont du système de règles et le suspend momentanément, le temps d'en modifier une et d'ouvrir ainsi de nouvelles branches, aussi cohérentes mais hétérogènes à la géométrie euclidienne.

Une machine pourrait-elle faire preuve d'une telle réflexivité ? On voit bien comment programmer *après-coup* une machine qui « change de règles » en passant d'un système d'axiomes à un autre, mais on ne voit pas comment programmer une capacité à reculer *d'un pas supplémentaire* et changer les règles qui président à ce changement de règles. Comment définir des règles qui présideraient à tout changement de règles ? Une machine peut répéter après-coup un geste comme celui de Lobatchewsky, mais non l'anticiper. Cela s'explique par le fait que le moment du changement de règles, pour lequel Bachelard emploie tout un vocabulaire révolutionnaire, implique une suspension partielle des règles. Il se produit une

539 Gaston Bachelard, *op. cit.*, p. 7.
540 Roger Caillois, *op. cit.*, p. 83.
541 Bachelard lui-même a un geste similaire lorsqu'il s'emploie à étudier la logique du rêve et de la rêverie. *La* logique devient *une* logique parmi d'autres. Voir la thèse de Julien Lamy, *Le pluralisme cohérent de la philosophie de Gaston Bachelard*, thèse de doctorat non publiée, soutenue en 2014 à l'Université de Lyon 3, sous la direction Jean-Jacques Wunenburgere.

mutation des principes, qui les transforme en un point précis, sans qu'il soit possible d'anticiper sur cette transformation : pourquoi ce principe et non un autre ? Et pourquoi le changer *de cette façon* ?

3.1.5. À quelles règles s'en tenir pendant la discussion des règles ?

Il est possible de concevoir un programme *pour chaque jeu particulier* (*ludus*), il est difficile d'imaginer quel type d'algorithme pourrait rendre compte *du jeu en général* (*paidia*). Cela implique de percevoir les signes qui manifestent le début d'un jeu spontané (une proposition explicite, un ton ironique, une posture… qui signalent, par exemple « qu'on joue à la marchande »), de pouvoir passer d'un jeu à un autre (« j'en ai assez, jouons au roi du silence ») et surtout d'inventer ou de s'adapter à de nouvelles règles (le roi du silence en se tenant sur un pied), voire à de nouveaux jeux. Comment formaliser ces mutations du corpus de règles qui, au cours du jeu, redonnent envie de jouer ? La langue française utilise l'expression *donner du jeu,* au sens où l'on dit qu'une porte a du jeu. Un peu d'espace entre les rouages facilite l'opération, mais trop d'espace la rend impossible. Il s'agit de trouver le bon écart qui produit l'aisance de mouvement, le bon niveau de difficulté pour que ce ne soit ni trop dur ni trop facile, ce qui dépend de l'interlocuteur (enfant, adulte, animal), du contexte (jeu en famille, entre amis, entre inconnus, jeu amoureux…), et maintient l'intérêt à participer[542]. Comment se décide et s'effectue l'évolution des règles au cours du jeu ? Est-il possible de rendre ce processus explicite ? Peut-on imaginer un jeu pendant lequel les participants feraient l'effort de rendre transparent chaque mutation de leur jeu ?

Dans un des dialogues d'introduction à *Vers une écologie de l'esprit,* Bateson met en scène un père et sa fille qui s'interrogent à ce sujet : « Est-ce qu'il y a des *règles* pour nos conversations[543] ? » Et si tel est le cas, en quoi consiste le fait de tricher pendant une conversation ? S'agit-il d'un jeu que l'on peut perdre ou gagner ? La fille demande au père :

542 Il faut préciser que la question de l'intérêt dépasse celle de la difficulté des règles. Le jeu sert toutes sortes de fonctions qui participent à nous y intéresser : faire connaissance, développer une complicité, départager les richesses, ouvrir des possibles, soulager le sujet en s'oubliant dans une activité absorbante, faire voyager par l'imaginaire. Caillois souligne que cette variété d'intérêts se combinent avec une gratuité du jeu, un goût fondamentalement désintéressé pour le jeu en lui-même. Le jeu doit préserver cette finalité en soi sinon il participe du travail. Un « mauvais joueur » bascule du côté de l'intérêt, se laisse emporter par le gain et transgresse ce principe du jeu comme finalité en soi.

543 Gregory Bateson, *Vers une écologie de l'esprit 1,* Paris, Éditions du Seuil, 1977, p. 44.

« Joues-tu *contre* moi ? » Une partie des règles de la conversation proviennent des idées abordées : « les idées avec lesquelles nous jouons introduisent certaines règles[544]. » Mais le fait de les suivre rigoureusement ne constituerait pas une conversation, cela reviendrait à « répéter les vieux clichés que tout le monde ânonne depuis des siècles[545]. » Aussi, pour « avoir des pensées originales et dire des choses nouvelles, nous devons briser toutes nos idées préconçues et en 'battre' les morceaux[546]. » Mais selon quelles règles effectuer ce mélange ? Cela ne risque-t-il pas de rendre fou ? De fait, les deux interlocuteurs ne cessent de tomber dans ce qu'ils appellent des « embrouillaminis », des moments où la confusion l'emporte.

> LE PERE : Attends un peu. Tu veux dire que les embrouillaminis c'est moi qui les provoque, en trichant avec les règles que d'ailleurs nous n'avons pas, ou, pour dire ça autrement, que nous pourrions avoir des règles qui, à condition que nous les observions, nous éviteraient les embrouillaminis.
>
> LA FILLE : C'est à ça que servent les règles d'un jeu !
>
> LE PERE : Oui, mais, si tu veux faire de nos conversations *ce* genre de jeu…, je préfère jouer à la canasta ; c'est plus drôle ;

Certaines règles implicites permettent de tenir une conversation, mais si on les suit de trop près, la conversation perd tout intérêt. Il faut prendre le risque de s'en écarter pour avoir un échange intéressant et que puisse émerger quelque chose d'inattendu. Sans cela, nous ne dirions jamais rien. « Si nous parlions toujours logiquement, sans tomber dans des embrouillaminis, nous ne pourrions jamais rien dire de nouveau[547]. » Et dans ce cas, autant s'adonner à un vrai jeu (la canasta), c'est-à-dire ne pas s'embarrasser à exprimer quelque chose (rester muets).

> LA FILLE : Bon, je vois ce que tu veux dire à propos de nous embrouillaminis… Ils nous font dire des choses nouvelles. Mais je pense à l'imprimeur. Il doit garder ses petites lettres bien classées, même s'il brise les phrases toutes faites. Et je me demande, à propos de nos embrouillaminis, si, pour ne pas devenir fou, il ne faut pas garder une sorte d'ordre dans les petits morceaux de notre pensée ?

544 *Ibid,* p. 45.
545 *Ibid,* p. 42.
546 *Ibid,* p. 43.
547 *Ibid.*

Une conversation oscille entre les « embrouillaminis » où il devient presque impossible de parler[548] et le respect de règles, qui ne font dire « que des phrases toutes faites ». Autrement dit, la suspension des règles rend muet, tout autant que le respect strict de la logique. Est-ce à dire que nous n'exprimons jamais rien ? Dans quelles conditions est-il possible de dire quelque chose ? Quel « ordre » faut-il garder « dans les petits morceaux de notre pensée » ?

Cette considération entraînant un nouvel « embrouillamini », les deux interlocuteurs en reviennent à leur propre conversation : si les règles sont introduites par les idées, d'où viennent ces idées ? Et d'où vient le passage d'une idée à une autre ?

> LE PERE : […] eh bien, oui, c'est moi qui fais les règles. Après tout, je n'ai nulle envie que nous devenions fous.
>
> LA FILLE : D'accord, mais est-ce que, des fois, en plus, tu les changes ?
>
> LE PERE : …. tu reviens à la charge. Oui, je les change constamment ; pas toutes, mais quelques-unes.

Les règles changent au fil de la conversation, sans que l'on puisse séparer le moment de la conversation et le moment où l'on définit les règles. La fille est scandalisée par ce changement des règles de la conversation *depuis l'intérieur* de la conversation (qui pourrait être qualifié de violation de la théorie des types) :

> LA FILLE : Tu pourrais peut-être me prévenir quand tu le fais !
>
> LE PERE : Encore ! Je voudrais bien, mais ce n'est pas comme ça que ça se passe. S'il s'agissait d'échecs ou de canasta, je pourrais t'indiquer les règles et, si nous le voulions, nous pourrions nous arrêter de jouer pour en discuter. Et puis, nous pourrions commencer un nouveau jeu avec de nouvelles règles. Mais à quelles règles s'en tenir entre les deux jeux ? Pendant la discussion des règles ?

En demandant à être prévenue quand il y a changement de règles, la fille propose une nouvelle règle quant au changement de règles. Il s'agirait de séparer nettement les niveaux de conversation : il y a la conversation, la conversation au sujet de la conversation, la conversation au sujet de la conversation qui porte sur la conversation, etc. La rigueur voudrait que soient

548 « LE PERE : Je sais. Nous voilà de nouveau dans un embrouillamini. Seulement, cette fois-ci, je ne vois aucun moyen d'en sortir. » *Ibid,* p. 45.

définies, pour chaque niveau, des règles appropriées, et que celles-ci soient rendues explicites. Mais il faudrait, pour pouvoir ajouter cette règle au sujet de la conversation, que la séparation existe déjà, que la conversation soit suspendue le temps de l'ajouter. En réclamant ainsi, la fille fait la même chose que ce qu'elle reproche à son père, elle change les règles depuis l'intérieur de la conversation, « sans prévenir ».

La réponse du père est qu'il est vain de vouloir séparer le moment du jeu du moment de définition des règles. Dans la mesure où il faut bien un dialogue pour s'entendre sur les règles du dialogue, on ne peut pas s'extraire absolument de la notion de conversation pour avoir une conversation sur la conversation. Ainsi, le père doit-il prévenir la fille quand il y a changement de règles au cours d'une discussion sur le fait de prévenir des changements de règles ? La conversation sur la conversation doit avoir lieu au sein d'une conversation, *il y a forcément fusion entre le niveau du dialogue et le niveau du métalogue* (dialogue sur le dialogue). Autrement dit, lorsqu'il y a réflexivité (ce qui porte sur la conversation *est une conversation*), la formalisation présuppose un niveau plus vaste où la théorie des types n'est pas respectée. On peut toujours remonter à ce niveau supérieur et le formaliser (définir les règles de la définition des règles), cela présuppose encore un niveau au-dessus (qui définit les règles qui définissent les règles qui définissent les règles) qui demeure implicite. Si on suspend une règle pour en discuter, il faut bien en discuter *selon d'autres règles* qui n'ont pas été suspendues. À cela s'ajoute que les règles peuvent être interdépendantes : les règles qu'on redéfinit affectent les règles qu'on garde pour les redéfinir. Enfin, certaines règles sont nécessaires à toute discussion : elles ne peuvent être suspendues pour en discuter puisque toute discussion les requiert.

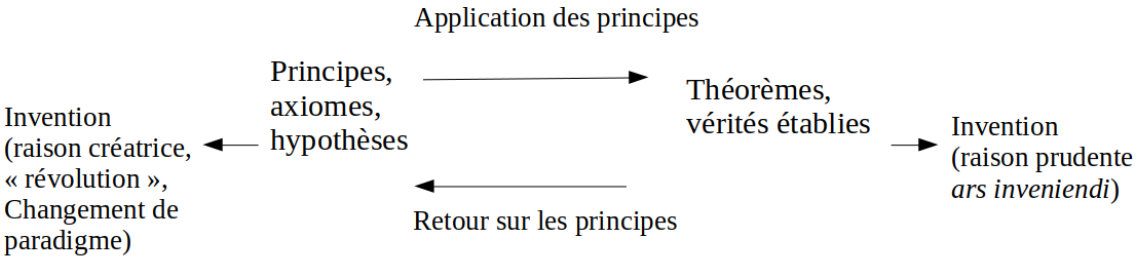
Dans le dialogue de Bateson, le père et sa fille se demandent à quel type de jeu correspond la conversation. Ils n'arrivent pas à la réduire à un jeu particulier définit par un système de règles (*ludus*), car une conversation au sujet du jeu (*ludus*), engendre plus de *ludus* qu'elle ne permet d'en formaliser. Vouloir formaliser une conversation semble aussi vain que de vouloir formaliser la *paidia,* et les deux personnages en arrivent à la conclusion qu'il est impossible d'en rendre compte :

> LA FILLE : Mais pourquoi les chatons et les chiots jouent-ils ?
>
> LE PERE : Je n'en sais rien[549].

549 *Ibid,* p. 47.

3.1.6. Les deux jambes de la pensée

Pour Bachelard il y a cohabitation de deux mouvements hétérogènes : celui d'une raison prudente, qui n'avance que de manière incrémentale, en capitalisant sur des résultats préexistants, et celui d'une raison créatrice qui remet en question les savoirs préétablis et avance par bifurcation. La première se donne pour tâche de réduire la pensée à des principes, c'est « une volonté acharnée d'atteindre le minimum d'hypothèses, le minimum d'éléments explicatifs ». Tandis que la deuxième apporte de l'hétérogène, elle « n'a pas peur de dépasser ce qui est l'exigence logique la plus stricte », il s'agit « de mettre le maximum de pensée dans le temps de la connaissance présente ». Il y aurait donc *deux types d'inventions* : celle de la raison prudente, qui capitalise sur les résultats préexistants, et celle qui passe par le retour sur les principes, à l'instar du geste opéré par Lobatchevski.



La raison prudente semble pouvoir être mécanisée tandis que la raison créatrice semble échapper à une description mécanique. Les programmes peuvent inventer, mais uniquement par capitalisation du savoir existant, et non par réflexivité. Ils peuvent proposer des formes inattendues, tout comme l'algèbre peut apporter de la nouveauté (*ars inveniendi*), mais sans opérer de suspension et de retour sur les principes.

Bachelard va plus loin. Les deux fonctions de la pensée, que Bachelard rebaptise « réduction » et « idonéisme » doivent s'entendre pour qu'un propos soit pertinent. Seule, la raison prudente n'avance guère. Seule, la raison créatrice se perd. Il faut qu'elles se rejoignent pour constituer une pensée.

> Réduction et idonéisme sont des fonctions indispensables pour la vie du nouvel esprit scientifique. Ces fonctions sont la systole et diastole qui doivent, sans fin, se succéder si nous voulons que la raison ait, comme il convient, une action de surveillance et une action d'invention, une action défensive et une action offensive[550].

La pensée adéquate selon Bachelard est conjonction des contraires, hiérogamie entre une recherche apollinienne de la cohérence et un goût dionysiaque pour la contestation. Il faut faire coexister le souci de la cohérence et de la consistence avec celui de la turbulence de l'agressivité – à la fois remettre en question les savoirs établis pour faire émerger l'invention et élaborer un édifice construit à partir de ce que l'invention produit.

La comparaison avec la systole et la diastole laisse penser qu'il s'agit de trouver le bon rythme, la bonne alternance entre réduction et idonéisme, entre moments apolliniens et moments dionysiaques. Mais aucune des deux fonctions ne peut fonctionner isolément : la réduction entraîne vers la tautologie tandis que l'idonéisme amène à la confusion. On retrouve le double écueil qui menaçait la conversation imaginée par Bateson : ou bien « [parler] toujours logiquement » et ne faire que ressasser de vieux clichés, ou bien, à trop changer les règles, ou réfléchir sur celles-ci, « tomber dans un embrouillamini » et se condamner au silence – toujours répéter la même chose ou ne rien dire du tout.

Cependant, il n'est pas si facile de distinguer les deux fonctions de la pensée. D'une part, tout discours possède sa loi de composition, il peut y avoir « réduction » à des principes, et cela vaut aussi pour les moments révolutionnaires : après coup, on peut donner les raisons du geste de Lobatchevsky, ou celles d'un « embrouillamini[551] ». D'autre part, chaque discours est toujours pris dans un devenir. À l'instar des axiomes et des principes de la géométrie euclidienne, tout corpus de règles qui sous-tend un discours finit par être amendé ou renversé pour donner lieu à un nouveau corpus de règles.

Ainsi, il n'y a pas à proprement parler d'alternance entre l'apollinien et le dyonisiaque mais entrelacement de ces deux tendances pourtant hétérogènes. Si l'apollinien peut donner des raisons pour tout, et même pour le dionysiaque, le dionysiaque introduit du devenir jusqu'à l'apollinien, c'est le principe de transformation des principes, la *paidia*, la fantaisie qui fait passer d'un ordre à l'autre. En d'autres termes, apollinien et dionysiaque, ou réduction et idonéisme, se conjuguent en une réorganisation permanente.

550 Gaston Bachelard, *op. cit.*, p. 28-29.

551 Cela vaut également pour le délire, en particulier lorsque celui-ci est dit « systématisé », mais aussi bien lorsqu'il ne l'est pas. « Dire n'importe quoi » est aussi une loi de composition du discours.

Chaque tendance pousse la pensée hors du temps dans une direction différente : l'apollinien tend vers l'éternel en mettant l'accent sur les structures qui peuvent s'abstraire du présent, alors que le dionysiaque entraîne vers « l'internel[552] » – un devenir pur qui nie toute structure. Seule leur union rend compte d'un discours en prise avec le souci de son temps.

3.1.7. La part rebelle : une objection à Turing

L'invention scientifique telle que Bachelard la décrit constitue une réponse à l'objection que Turing fait à Lovelace. Les machines sont un support d'invention (*ars inveniendi)*, et la combinatoire qu'elles opèrent ressemble à notre créativité, mais il leur manque une part dionysiaque, l'élément rebelle sur lequel Bachelard met l'accent. Aussi, les machines peuvent nous surprendre par les combinaisons qu'elles opèrent au sein de principes donnés, mais elles ne peuvent restituer la façon qu'a la pensée humaine de déjouer toute attente en jouant sur les principes eux-mêmes. Chaque manifestation de cette capacité à surprendre peut être décrite et reproduite après-coup mais elle n'a pas de mécanisme universel qui permettrait de la reproduire ou de l'anticiper. Priver la raison de cet élément rebelle reviendrait à la réduire à l'activité de la mémoire :

> On confond presque toujours l'action décisive de la raison avec le recours monotone aux certitudes de la mémoire. Ce qu'on sait bien, ce qu'on a expérimenté plusieurs fois, ce qu'on répète fidèlement, aisément, chaleureusement, donne une impression de cohérence objective et rationelle. Le rationalisme prend alors un petit goût scolaire[553].

Une raison trop prudente se cantonne aux champs déjà délimités et s'interdit de produire de la nouveauté. En ne procédant que du passé, la raison se prive d'avenir, elle ne fait que confirmer encore et encore ce que nous savons déjà. Bachelard se moque de l'image d'une raison « capitalisante » qui va avec l'idée d'un progrès linéaire, d'un « enrichissement monotone » où chaque chercheur vient ajouter sa pierre, « enrichir » le « patrimoine intellectuel » d'un seul et même édifice commun, sans imprudence et sans risque. Pour retrouver de *l'intérêt* (autrement dit, du sens), il faut que la raison *s'investisse* hors de ce qu'elle sait déjà. Bachelard plaide pour

552 Le terme est emprunté par Gilles Deleuze à Charles Peguy. Il le compare à « l'intempestif » de Nietzsche. Gilles Deleuze, Félix Guattari, *Qu'est ce que la philosophie,* Paris, Éditions de minuit, 1991.
553 Gaston Bachelard, *op. cit.*, p. 7.

« rendre à la raison humaine sa fonction de turbulence et d'agressivité[554] », de façon à « tourner […] le rationalisme du passé de l'esprit à l'avenir de l'esprit, du souvenir à la tentative, de l'élémentaire au complexe[555] [...] »

Croire que la raison n'a pas d'histoire et qu'il faut l'élaborer pas à pas, sans révolution, revient à l'empêcher de se renouveler, à l'enfermer dans une mécanique de répétition et de thésaurisation vaine. « Les connaissances longuement amassées, patiemment juxtaposées, avaricieusement conservées, sont suspectes. Elles portent le mauvais signe de la prudence, du conformisme, de la constance, de la lenteur[556]. » Bachelard veut dépasser l'image d'une raison figée, hors du temps, pour laquelle rien d'autre ne peut avoir lieu que la révélation progressive de ce qu'elle a toujours été. Au contraire il faut redonner du mouvement à la raison, lui restituer sa jeunesse, opter pour ce qu'il qualifie de *surrationalisme*, en référence à la subversivité des surréalistes.

> Il suffit […] de se débarasser de cet idéal d'identification pour que le mouvement s'empare tout à coup des dialectiques rationnelles. Alors, au rationalisme fermé fait place le rationalisme ouvert. La raison heureusement inachevée ne peut plus s'endormir dans une tradition ; elle ne peut plus compter sur la mémoire pour réciter ses tautologies. Sans cesse, il lui faut prouver et s'éprouver. Elle est en lutte avec les autres, mais d'abord avec elle-même. Cette fois, elle a quelques garanties d'être incisive et jeune[557].

Cela implique de se résigner au fait que chaque époque de la pensée sera un jour révolue. Il y a des époques qui ne reviennent plus, il y a des penseurs uniques. Aussi, Bachelard en appelle à « singulariser les diverses philosophies rationalistes, à réindividualiser la raison[558]. » Il s'agit d'affirmer que la raison participe de l'histoire et de ses polémiques. Bachelard évoque « la dialectique instituée au niveau des notions particulières, *a posteriori,* après que le hasard ou l'histoire ont apporté une notion qui reste, par cela même, contingente[559]. »

En faisant référence au hasard, à l'histoire, à la contingence, Bachelard s'élève contre une image déterministe ou déterminée de la raison, et s'attaque plus particulièrement aux logiciens et mathématiciens souscrivent au projet formaliste : « les logiciens et les formalistes ont, tout au contraire, déréalisé, dépsychologisé, la nouvelle conquête spirituelle[560]. » Il

554 *Ibid.*
555 *Ibid.*
556 *Ibid,* p. 11
557 *Ibid,* p. 12.
558 *Ibid,* p. 9
559 *Ibid.*
560 *Ibid.*

préconise de prendre le contrepied de leur méthode, de ne pas se bander les yeux pour ne compter que sur la forme car, « après cette œuvre de mises en formes bien vidées de toute pensée, après cette besogne de sous-réalisme acharné, l'esprit n'est pas devenu plus alerte et plus vivant, mais plus las et plus désenchanté[561]. » Il faudrait au contraire refaire signifier les symboles. Le « devoir du surrationalisme » est de reprendre les formes et « de les remplir psychologiquement, de les remettre en mouvement et en vie[562]. » En d'autres termes, il fait un plaidoyer pour l'intuition au sens où nous l'avons défini (voir section 2.6.5. Récapitulatif : une définition de l'intuition).

L'image de la raison proposée par Bachelard vient au secours de l'objection de Lovelace et apporte une réponse à l'ironie de Turing. Selon Turing, les machines nous surprennent car nous ne faisons pas assez de calcul pour prévoir leur comportement. Sur ce point, il semble être d'accord avec Lovelace pour qui les machines ne peuvent « être à l'origine » d'une idée (« to originate anything »). Elles ne font que dérouler une procédure, sans qu'il y ait émergence de nouveauté. L'objection de Lovelace souligne une diffférence avec les humains qui seraient, eux, capables de produire de la nouveauté. Mais pour Turing, la surprise que provoque un humain pourrait être interprétée de la même façon que celle créée par une machine : peut-être qu'un humain ne semble créatif que parce que nous ne prenons pas le temps de prévoir ses pensées. Turing insinue que nous pourrions, en droit, ne jamais être supris par une idée. Ce n'est que par *manque de moyens* que nous sommes surpris, parce que nous ne prenons pas le temps de suivre le même cheminement de pensée. Nous parlons de création alors qu'il y a seulement excès : nous sommes *dépassés* par les opérations qu'effectue quelqu'un de réputé créatif. Aussi, ces opérations peuvent être complètement déterminées et surprendre ceux qui ne les ont pas effectuées, donnant l'impression qu'il y a eu invention, alors que nous ne sommes pas plus « à l'origine » d'idées que les machines. Autrement dit, Turing semble sous-entendre que les opérations de l'esprit sont calculables même quand elles nous surprennent, et qu'elles pourraient toujours être rendues explicites par une série d'opérations sur des symboles.

En mentionnant le hasard et en inscrivant la raison dans l'histoire, Bachelard propose une interprétation opposée de la créativité. Il fait allusion à la pensée aveugle, ces « mises en formes bien vidées de toute pensée », et lui oppose une faculté à ajouter du jeu dans l'agencement des symboles. Ce jeu intervient à **deux niveaux** : un premier niveau consistant en une **variation dans les règles et principes** qui encadrent nos réflexions (à l'instar du changement d'axiome opéré par Lobatchewski), et un deuxième niveau consistant en une

561 *Ibid.*
562 *Ibid.*

variation dans la façon dont ces symboles sont interprétés – dans ce que la pensée y « voit » ou les enjeux qu'elle y « investit ». Ce sont deux manières d'utiliser l'intuition telle que nous l'avons définie : par un retour sur les principes qui amène à laisser la possibilité de surgissement de nouveaux principes, ainsi que par un renouvellement de l'attribution du sens (de l'interprétation) – l'un et l'autre étant liés puisque l'attribution du sens passe par la perception des principes. En ce sens, pour Bachelard, nous sommes bien, via l'intuition, « à l'origine de quelque chose » qui s'ajoute à l'opération de symboles – la mutation des règles d'opération ainsi que le point de vue porté sur les symboles (la façon dont ils sont investis). Il y a là un devenir qui ne paraît pas avoir de principe, qui n'est pas calculable. L'image de la raison « surrationaliste » que propose Bachelard nous permet de souligner ce qui manque à l'article de Turing : l'existence d'une part contingente de la raison sans laquelle il n'y aurait pas de créativité. Il y a de la nouveauté, et celle-ci ne saurait être restituée par un algorithme, puisque la nouveauté s'introduit par un jeu sur l'algorithme et sur la manière dont il est interprété.

C'est notamment ce qui nous permet de *traverser une crise.* En mars 2020, à l'occasion de la pandémie de Covid-19, nos comportements changent brutalement : confinements, distanciation, achats compulsifs, télétravail… L'état d'exception suscité par l'irruption du virus engendre aussitôt l'invention de nouvelles règles de comportement et la réinterprétation d'anciennes règles – par exemple, ce que signifie « être présent au travail ». Les algorithmes – « apprenants » ou non – , n'ont pas cette capacité à moduler leur corpus de règles pour garder de la pertinence malgré le changement de situation. Déroutés par ce qui, du point de vue antérieur à la crise, sont des comportements aberrants, les algorithmes de prédiction des ventes, de recommandation de produits, de prévision des stocks, deviennent inutilisables[563]. Leur comportement apparaît aberrant par rapport à ce qu'exige le changement de situation. Ils ont été incapables de la *normativité* dont a fait preuve l'humain[564].

Cette nouveauté de la normativité ne saurait être restituée par un algorithme, puisqu'elle s'introduit par un jeu sur l'algorithme et sur la manière dont il est interprété. Pour autant, est-ce que ces arguments permettent d'invalider le projet d'intelligence artificielle ? Cette raison qui possède une part rebelle, ne serait-il pas possible de la reproduire sous la forme d'une machine qui fait une place au hasard ? Est-ce que l'introduction de hasard peut suffire pour *donner du jeu* à la machine et lui permettre de passer pour créative ?

563 Will Douglas Heaven, « Our weird behavior during the pandemic is messing with AI models », *op. cit.*

564 La normativité est un des thèmes de prédilection de Georges Canguilhem. Elle ne concerne pas seulement les humains mais l'ensemble des organismes vivants et se caractérise par leur capacité, en situation de crise (par exemple, la maladie), de créer de nouvelles normes. Georges Canguilhem, *Le normal et le pathologique*, Paris, PUF, 2013.

3.2. Donner du jeu : le hasard

3.2.1. « L'injection de hasard »

Dans le septième paragraphe de la proposition de Dartmouth, intitulé « le hasard et la créativité[565] », les auteurs estiment que « la différence entre la pensée créative et la pensée compétente sans imagination repose dans l'injection de hasard » – tout en gardant une certaine prudence, puisqu'ils qualifient l'idée d'« assez séduisante mais clairement incomplète[566] ». Ils précisent que « le hasard doit être guidé par l'intuition pour être efficace. En d'autres termes, l'estimation éclairée ou le pressentiment inclut un hasard contrôlé dans une pensée par ailleurs ordonnée[567]. » Aussi, le hasard qu'il faut « injecter » pour simuler la créativité n'est pas un hasard injecté *au hasard.* Il est « guidé » par une faculté, l'intuition, qui aurait donc la capacité de savoir en avance dans quelle direction se trouve la solution. C'est uniquement après que cette direction ait été devinée que l'esprit se permettrait de recourir au hasard pour trouver une solution, de la même façon que, au moment de chercher un objet dans une maison, on se donnera des directions préalables. Si la salière a disparu, on la cherchera d'abord dans la cuisine, pour éventuellement y chercher au hasard, mais sans aller dans la salle de bain. Cette conception de la créativité pose problème. Comment se pourrait-il que ce qui est recherché, et qu'on ne connaît pas encore, soit préalablement localisé par l'intuition, comme s'il connaissait *déjà* l'objet de la quête ?

Les données du problème pourraient fournir une idée de la solution, tout comme on sait que la salière est probablement dans la cuisine. À cela s'ajoute que l'expérience permet de guider la recherche. C'est le grand nombre de fois où j'ai vu la salière dans la cuisine qui me permettra de « deviner » qu'elle y est probablement. Mais d'où peut venir « l'estimation éclairée ou le pressentiment » d'une solution à un problème nouveau, en particulier si ce problème implique un changement de paradigme ? Comment puis-je avoir l'idée de suspendre un des axiomes de la géométrie euclidienne ? Comment puis-je savoir par avance ce qu'il est

565 « Randomness and creativity », in John McCarthy, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon, « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 », *op. cit.*

566 « A fairly attractive and yet clearly incomplete conjecture is that the difference between creative thinking and unimaginative competent thinking lies in the injection of a some randomness. » *Ibid.* Nous traduisons.

567 « The randomness must be guided by intuition to be efficient. In other words, the educated guess or the hunch include controlled randomness in otherwise orderly thinking. » *Ibid.* Nous traduisons.

pertinent d'explorer ? Il existe de nombreux exemples d'inventions où la bonne idée vient de domaines hétérogènes aux données du problème. C'est l'argument que donne Fodor pour montrer que la découverte scientifique ne peut pas être une fonction spécifique (ou « modulaire ») de l'esprit[568]. Comme nous ne savons pas quelles causes régissent les phénomènes, « nous devons être prêts à changer d'avis sur ce qui est pertinent du point de vue de la confirmation à mesure que nos théories scientifiques changent[569]. » Cela rend impossible de restreindre la recherche de la solution à un domaine spécifique : « tout le savoir du scientifique est pertinent en principe pour déterminer quelles nouvelles croyances il devrait adopter[570] ». En d'autres termes, est-ce qu'il ne faudrait pas que la direction de recherche, l'intuition qui « guide » le hasard, *contienne elle-même* du hasard, pour prendre en compte la faible possibilité que la salière soit dans la chambre ou la salle de bain ?

D'après les auteurs de l'appel à projet de Dartmouth, « l'estimation éclairée ou le pressentiment inclut un hasard contrôlé ». Mais il y a une ambigüité sur le sens du mot « inclure » : cette phrase est-elle une reformulation de la phrase précédente, c'est-à-dire que « l'estimation éclairée » est *ce qui cadre* le travail du hasard (diriger les recherches vers la salle de bain) ou bien une idée supplémentaire affirmant que « l'estimation éclairée » est *elle-même hasardeuse* (et si on regardait vers la cage d'escalier) ? Mais ce hasard supplémentaire, qui s'applique à ce qui guide le premier hasard, *par quoi est-il lui-même guidé* ?

Les auteurs prennent soin de préciser que, quel que soit le niveau où intervient le hasard (la recherche ou la direction de recherche), il s'agit toujours d'un hasard contenu. Par trois fois, ils renient l'idée que la pensée puisse être hors de contrôle : d'abord ils écrivent que le hasard est « guidé » par l'intuition, puisqu'il est « un hasard contrôlé » et enfin ils précisent que la pensée est « par ailleurs ordonnée ». Les auteurs admettent que leur modèle du hasard est « incomplet » mais ils s'opposent d'avance à l'hypothèse, par ailleurs plausible, d'une créativité laissant la place à de pures divagations. Ils le rejettent par trois fois, avec trois nuances importantes. Premièrement, le hasard est circonscrit : *il intervient à certains endroits, mais pas*

568 Le même point de vue se trouve chez les tenants de la *Gestalt theory,* chez Von Neumann, ou encore chez ceux qui s'intéressent à la résilience du cerveau et des organismes en général. « Organisms are stable as *unities* precisely because their organization is *not* fixed into any one rigid structure ; they are open, and thus equipped to surmount even a traumatic loss of functions in some cases. » David Bates, « Automaticity, Plasticity, and the Deviant Origins of Artificial Intelligence », in Bates et Bassiri (eds.), Plasticity and Pathology : On the Formation of the Neural Subject, New York, Fordham University Press, 2015, p. 207. voir également David Bates, « Creating Insight : Gestalt Theory and the Early Computer » in Jessica Riskin (ed.), *Genesis Redux, Essays in the History and Philosophy of Artificial Life,* Chicago, The University of Chicago Press, 2007, p. 238-239.

569 Jerry Fodor, *La modularité de l'esprit : essai sur la psychologie des facultés,* Paris, Les Éditions de Minuit, 1986, p. 137.

570 *Ibid.*

ailleurs, puisqu'ailleurs la pensée est « bien ordonnée ». Ensuite, le hasard est tenu en laisse – *il y a des possibilités admises et d'autres qui sont exclues* – puisqu'il est « contrôlé ». Enfin, *le sujet a déjà une idée de ce qu'il va produire,* ou au moins une « estimation éclairée », un « pressentiment », et il l'impose comme cadre : l'exercice du hasard est « guidé » par l'intuition.

Dans la partie de l'appel à projet de Dartmouth rédigée par Nathanael Rochester, celui-ci reprend l'idée d'une machine qui « montrerait de l'originalité dans ses réponses aux problèmes[571] » en intégrant le hasard (*randomness*). Le hasard, écrit-il, permet de passer outre la myopie et les préjugés du programmeur (« the shortsightedness and prejudices of the programmer »), autrement dit d'explorer toute la gamme des possibles. Surtout, il semble indispensable au traitement des problèmes nouveaux, ces problèmes pour lesquels « aucun individu de la culture n'a de solution et qui a résisté aux efforts[572] » et qui requièrent « de faire quelque chose qui est déraisonnable ou inattendu au regard de l'héritage des connaissances accumulées par la culture[573] ». Pour dévier des connaissances établies et trouver cette action il faut un pressentiment (*hunch*), quelque chose « d'inattendu, de pas tout à fait raisonnable[574] ». Un peu de hasard (« just a little randomness »), pourrait-il fournir ce décalage ?

> Si, par exemple, on écrivait un programme qui, une fois toutes les 10 000 instructions, prenait un nombre aléatoire et l'exécutait comme instruction le résultat serait probablement le chaos. Après uns certaine dose de chaos la machine essayerait probablement quelque chose d'impossible [*forbidden*] ou exécuterait une instruction d'arrêt et l'expérimentation s'arrêterait là[575].

Autrement dit, le hasard seul ne peut fonctionner : « ça ne marchera pas d'introduire le hasard sans utiliser l'intuition [forsight][576] ». Mais comment intégrer ce rôle de l'intuition ? Sans donner de réponse directe, Rochester propose deux pistes de recherche : étudier le cerveau pour

---

571 « how can I make a machine which will exhibit originality in its solution of problems? », Nathanael Rochester, « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 », *op. cit.*

572 « Now consider a problem for which no individual in the culture has a solution and which has resisted efforts at solution. » *Ibid.* Nous traduisons.

573 « In order to solve this problem the individual will have to do something which is unreasonable or unexpected as judged by the heritage of wisdom accumulated by the culture. » *Ibid.* Nous traduisons.

574 « The individual needs a hunch, something unexpected but not altogether reasonable. » *Ibid.* Nous traduisons.

575 « If, for example, one wrote a program so that once in every 10,000 steps the calculator generated a random number and executed it as an instruction the result would probably be chaos. Then after a certain amount of chaos the machine would probably try something forbidden or execute a stop instruction and the experiment would be over. » *Ibid.* Nous traduisons.

576 « it will not do to introduce randomness without using forsight. » *Ibid.* Nous traduisons.

copier son fonctionnement, ou bien s'essayer à résoudre des problèmes réels exigeant de l'originalité. Il ne donne aucune définition de l'intuition et conclut le texte par des considérations sur la différence entre les erreurs faites par un programme, et celles faites par un humain, qui « parfois ont presque du sens », lorsque l'humain est endormi, ivre, ou fiévreux[577]. Ce qui l'amène à se demander si

> le mécanisme du cerveau est tel qu'une petite erreur de raisonnement introduit, exactement de la bonne manière, du hasard. Peut-être que le mécanisme contrôlant l'ordre du comportement guide le facteur hasard de manière à améliorer l'efficacité du processus imaginatif par rapport à du pur hasard[578].

À nouveau, le hasard n'intervient qu'à des moments bien précis, « exactement de la bonne manière », et dans un dosage limité – lors d'une « petite erreur de raisonnement » – restant sous contrôle d'une intuition dont la notion reste mystérieuse.

En 1948, quelques années avant Rochester, Turing était allé plus loin en s'interrogeant sur la possibilité et l'intérêt de fabriquer des « machines partiellement aléatoires » (« partially random and apparently partially random machines »), voire des « machines non organisées » (« unorganised machines »), de manière à imiter le cortex en tant que machine « non organisée » capable de s'auto-modifier à la faveurs d'« interférences » et se paramétrant par les stimulis de l'éducation (« pleasure-pains systems »)[579]. Mais chaque effort pour *décrire* une telle machine revient également à circonscrire le hasard. Telle qu'elle est décrite dans l'article de Turing, et en particulier dans celui de Rochester, la place du hasard dans le processus de créativité ressemble à celle du jet de dé lors d'un jeu de plateau. On ne tire pas les dés à tout bout de champ puisque par ailleurs, le jeu est « bien ordonné ». Le dé a une gamme de possibilités prédéfinie par ses faces : on peut tirer un ou cinq, mais pas zéro ou sept, ni une lettre ou une image. Et si le jet de dé apporte une incertitude sur l'issue du jeu, c'est une incertitude restreinte. Elle ne fait varier que la durée du jeu, l'identité du gagnant et le montant de son gain. S'il peut y avoir des surprises, celles-ci ne renversent jamais l'ordre établi du jeu, c'est une créativité tenue en laisse, avec une **prééminence nette du contrôle sur le hasard.**

577 « On the other hand human errors in speech are apt to result in statements which almost makesense (consider someone who is almost asleep, slightly drunk, or slightly feverish). » *Ibid.*

578 « Perhaps the mechanism of the brain is such that a slight error in reasoning introduces randomness in just the right way. Perhaps the mechanism that controls serial order in behavior guides the random factor so as to improve the efficiency of imaginative processes over pure randomness. » *Ibid.*

579 Alan Turing, « Intelligent Machinery », *in* Jack Copeland, *The Essential Turing, op. cit.,* p. 410-432.

3.2.2. La part mobile

Dans le texte de fiction « La loterie à Babylone[580] », José-Luis Borges remarque, au sujet des loteries : « ne conviendrait-il pas que le hasard intervînt dans toutes les étapes du tirage et non point dans une seule ? » En effet, pourquoi le hasard s'appliquerait-il seulement à la sélection du gagnant ? Pourquoi ne pas le faire intervenir sur le nombre de gagnants, le montant du gain, ou le délai de réception ? Et pourquoi ne pas faire intervenir le hasard sur le lieu de l'intervention du hasard ? Il pourrait, lors de la première partie, faire varier le nombre de gagnants, puis, à la suivante, faire varier le montant du gain… Le hasard pourrait également s'appliquer au rythme de changement des règles : deux parties successives ont les mêmes règles, puis trois règles changent à la suivante, puis deux règles, etc. Autrement dit, le hasard pourrait intervenir sur les règles qui président à ces changements de règles. Si la part aléatoire est délimitée, le jeu reste cantonné à un *ludus* trop simple. Mais si la part aléatoire est « libre », si elle peut s'appliquer à toutes les règles sans restriction, alors cette part n'est pas formalisable : à chaque fois qu'on la formalise, on ajoute un niveau de règles auquel l'aléatoire pourrait s'appliquer, à l'infini. Comment concevoir un tel jeu sans sombrer dans la confusion ?

Dans un chapitre de *Logique du sens* consacré aux jeux, Gilles Deleuze mentionne Lewis Carroll qui « invente des jeux, ou transforme les règles de jeux connus (tennis, croquet) » et surtout

> invoque une sorte de jeu idéal dont il est difficile à première vue de trouver le sens et la fonction : ainsi dans *Alice* la course à la Caucus, où l'on part quand on veut et où l'on s'arrête à son gré ; et la partie de croquet, où les boules sont des hérissons, les maillets des flamants roses, les arceaux des soldats qui ne cessent de se déplacer d'un bout à l'autre de la partie. Ces jeux ont ceci de commun : ils sont très mouvants, ils semblent n'avoir aucune règle précise et ne comporter ni vainqueur ni vaincu[581].

580 Jorge Luis Borges, « La loterie à Babylone », in *Fictions,* traduit de l'espagnol par P. Verdevoye, Ibarra et R. Caillois, Paris, Gallimard / Folio, 1999.
581 Gilles Deleuze, *Logique du sens,* Paris, Éditions de Minuit, 1969, p. 74.

Alors que « nos jeux connus répondent à un certain nombre de principes, qui peuvent faire l'objet d'une théorie[582] », les jeux de Carroll jouent sur les principes, et déjouent toute théorie. Nos jeux sont des « jeux partiels » par rapport à un jeu pur, jeu idéal où la part aléatoire circule sans entraves. Mais est-ce qu'un tel jeu pur est seulement concevable ? Est-ce qu'on peut encore 'jouer' à de tels jeux où il n'y a ni gagnant, ni perdant, et où les changements constants font que l'on ne sait plus *à quoi* on joue ?

Si les auteurs de la conjecture de Dartmouth donnent une image si pauvre du hasard à « injecter » dans un programme pour rendre compte de la créativité, c'est sans doute pour éviter la confusion que provoque la perspective d'un véritable hasard, une « injection » mobile, qui peut intervenir à n'importe quel moment du processus, d'une façon non prédéfinie. Autrement dit, les auteurs postulent une part fixe (un jeu partiel) parce qu'ils ne sauraient que faire d'une part mobile (jeu pur). S'il existe une part mobile, et que celle-ci est impossible à formaliser, cela invaliderait la conjecture de départ selon laquelle tous les aspects de l'intelligence peuvent être décrits avec une précision telle qu'une machine peut les simuler. Il faut une part fixe, non parce que c'est la meilleure solution, mais car c'est la seule qu'on puisse représenter avec précision. Mis à part ces réticences pratiques, qu'est-ce qui pourrait s'opposer à une notion de la créativité comme « jeu pur » sujet à une « part mobile » de hasard, tel que le décrit Deleuze ? Une telle notion de la créativité ne semble-t-elle pas plus pertinente que celle d'un hasard tenu en laisse par l'intuition ?

3.2.3. Un hasard épouvantable

Dans *Logique du pire,* Clément Rosset distingue quatre niveaux de hasard[583] :

Il y a **le sort**, c'est-à-dire la « responsabilité d'une série causale heureuse ou malheureuse ». L'événement hasardeux est rapporté à une cause – la fortune, ou la nécessité, éventuellement sous une forme déifiée – ce qui en fait moins du hasard que son exact contraire, le destin, dans la mesure où sont présupposées l'idée d'un enchaînement causal et l'idée d'une finalité.

582 Deleuze en donne les éléments – « les règles catégoriques préexistantes, les hypothèses distribuantes, les distributions fixes et numériquement distinctes, les résultats conséquents » – et les détaille page 74.
583 Clément Rosset, *Logique du pire,* Paris, PUF, 1993, p. 73-76.

Un deuxième niveau de hasard est **la rencontre** ou « point d'intersection entre deux ou plusieurs séries causales ». Ici, « le fortuit s'est déplacé de l'ensemble d'un enchaînement au caractère imprévisible de la rencontre, en certains points, de certains enchaînements. » C'est le hasard selon Cournot comme rencontre de deux séries causales indépendantes. Si la *coïncidence* est qualifiée d'hasardeuse, on présuppose encore des séries causales constituées à partir de laquelle elle a lieu.

Un troisième niveau de hasard est **la contingence, «** dérivée elle aussi de l'idée de simultanéité (*cum-tangere*) », elle « ne désigne plus le fait hasardeux à la faveur duquel deux séries coïncident, mais le principe général d'imprévibilité qui est attaché à de telles rencontres. » C'est l'idée qu'une partie des événements n'est pas nécessaire : « si tout n'est pas prévisible, c'est – peut-être – que tout n'est pas nécessaire ; il pourrait donc y avoir de la non-nécessité, qu'on appellera contingence. » Cette dernière présuppose donc l'idée de nécessité : la nécessité serait l'état « naturel » du cours des choses et pourrait être suspendue pour certaines catégories d'événements (émergence de la vie, de la conscience, décision libre d'un humain, délire…). C'est ce troisième type de hasard qui est mentionné par la proposition de Dartmouth pour rendre compte de la créativité. Les auteurs font l'hypothèse qu'il y aurait une contingence locale sur fond de nécessité. Ainsi, lorsqu'il y a « injection de hasard », la nécessité ou l'ordre de cette pensée « par ailleurs bien ordonnée » est suspendue temporairement et de manière localisée (« contrôlée » et « guidée »).

Clément Rosset appelle le quatrième et dernier niveau de hasard – **le hasard,** puisqu'il est pour lui le seul véritable hasard. Le mot aurait lui-même une origine hasardeuse, du nom d'un château en Syrie où aurait peut-être été inventé le jeu de dé. « Hasard » désigne la face du dé qui porte le nombre six, « jeter hasard » signifiant qu'on a obtenu le six. Plus tard, le mot prend la signification « de risque, de péril, de situation se dérobant à toute possibilité de *contrôle*[584] ; ».

Les trois premiers hasards correspondent à « un hasard d'après la nécessité (et les séries causales) », ce qui pour Clement Rosset n'est pas vraiment du hasard. Un véritable hasard est « un hasard d'avant la nécessité.» Ce qui revient au « vieux problème de savoir si le désordre ne peut se concevoir qu'à partir de l'ordre (thèse de Bergson), ou si l'on peut parler, avec Lucrèce, de désordre et de hasards originels[585] ».

Il a fallu mentionner les trois premiers « hasards » bien qu'ils n'en soient pas à proprement parler, car c'est eux qui sont le plus souvent désignés par l'usage du mot. En effet,

584 *Ibid.*
585 Clément Rosset, *op. cit.,* p. 72.

le quatrième – et seul véritable – « hasard » est généralement soigneusement évité. C'est que l'idée d'un désordre originel, « antérieur à toute idée d'ordre ou de désordre », est à la fois difficile à penser et suscite l'épouvante.

> Lorsque Pascal parle de hasard, ce n'est pas l'imprévisibilité des rencontres qui est en question, ni la possibilité philosophique de la non-nécessité, mais plutôt l'intuition d'un manque à penser, d'un blanc, d'un silence, antérieurs à toute possibilité de rencontre (qui suppose un monde constitué) comme à toute possibilité de pensée (qui suppose la création de l'homme). En ce sens, « hasard » désigne, chez Pascal, très précisément l'enfer[586].

Ce hasard est une notion ardue : c'est « l'intuition d'un manque à penser ». Par rapport aux trois autres notions de hasard, c'est la seule qui ne s'appuye sur aucun présupposé. Il est, d'une certaine façon, impensable, il « désigne l'acte même de la négation, sans référence précise à ce qu'il nie. Ignorance originelle, appellée à ne nier qu'accessoirement, et après coup, tout ce qui pourrait se constituer comme pensée[587]. » Comme on ne peut presque rien en dire, sa définition ne peut être que négative : ce n'est *ni* le sort, *ni* la rencontre, *ni* la contingence. Ce qui le caractérise, c'est qu'il n'y a *aucune intervention* : ni d'une divinité, ni de la causalité, ni d'une contingence qui introduirait des exceptions à la causalité – « interdiction de tout recours extérieur, s'appelât-il chance, destin, providence ou fatalité[588]. » Ainsi, il est exact que « la matière inerte reçoit du hasard ce qu'on appelle la vie, le mouvement et les différentes formes d'ordre », bien que le terme « recevoir » soit impropre « puisqu'il suppose l'existence de deux instances différentes, dont l'une, le hasard, imprimerait vie (et nature) à l'autre, la matière[589]. » Une pensée à la hauteur du hasard conçoit « l'aptitude de la matière à s'organiser spontanément » sans qu'il y ait impression d'une forme extérieure.

Si les auteurs de la proposition de Dartmouth formulent leur hypothèse selon le paradigme de l'intervention (intervention d'une dose de contingence dans une pensée « par ailleurs ordonnée »), c'est qu'ils ne peuvent pas faire autrement : leur point de départ est une description de l'intelligence à laquelle « manque » une faculté, la créativité. Ce « manque » est interprété comme un « manque » de hasard qu'il faut par conséquent « ajouter » au modèle. Or, s'il faut « l'ajouter » cela ne peut être un hasard radical puisqu'il y a « intervention ». À partir du moment où il arrive après-coup, c'est toujours un hasard de pacotille. Autrement dit, aucun

586 Clément Rosset, *op. cit.*, p. 77.
587 *Ibid*, p. 78.
588 *Ibid*, p. 77.
589 *Ibid*, p. 85.

« manque » de hasard ne saurait être « comblé » par l'ajout de quelque chose d'autre (d'une « dose » de hasard). Cela ne fait que réitérer le paradigme de l'intervention et passer à côté du hasard radical.

3.2.4. « L'aptitude de la matière à s'organiser spontanément »

La prise en considération de la pertinence du hasard pour décrire ce qu'est la « part rebelle » ou l'élément dionysiaque laisse penser que c'est elle qui a la prééminence. Il ne s'agit pas, comme le défend Bachelard, d'un équilibre entre organisation et révolution. De la même façon que selon Caillois le *ludus* procède de la *paidia* et que selon Rosset les stabilités transitoires que nous prenons pour des « natures » naissent du hasard radical, les structures de pensées surgissent sans raison. Pour le dire autrement, la pensée commence par l'erreur, ou plus précisément par l'errance[590]. Le point de vue des auteurs de la proposition de Dartmouth est inversé : ce n'est pas le hasard qui intervient au sein de l'ordre, mais différents ordres qui se constituent et se succèdent par hasard. Il n'y a pas « un hasard contrôlé dans une pensée par ailleurs ordonnée » mais émergence d'automatismes depuis un fond sans nécessité. Les auteurs de la proposition de Dartmouth veulent une pincée de hasard pour modéliser la créativité, mais *ce hasard là* est impossible à modéliser – c'est de là que naissent les modèles.

Si l'on adopte ce point de vue, la part de hasard que manifeste notre créativité ne serait pas un privilège de l'humain ou de la vie puisqu'elle participe de « l'aptitude de la matière à s'organiser spontanément[591] ». La notion de créativité ne permet plus de distinguer l'humain de la matière, ce qui rejoint le point de vue de Turing, mais pour des raisons inverses. Selon lui, si l'humain et la machine peuvent nous surprendre, c'est que nous ne prenons pas le temps de faire les opérations qu'ils effectuent – celles-ci sont déterminées, mais trop complexes pour que nous ayons les moyens de les calculer. Du point de vue du hasard, cette capacité commune à la surprise peut s'expliquer par le fait que les humains comme les machines participent du même réel non déterminé. Cela revient à défendre la même cosmologie « participative » que le projet

590 « On pourrait même dire que toute pensée véritable est une erreur, parce qu'elle se déprend de la connaissance existante » David Bates, « Automatisation et erreur », in Bernard Stiegler (ed.), *La verité du numérique : Recherche et enseignement supérieur à l'ère des technologies numériques,* Paris, FYP Éditions, 2018, 29-40.

591 « Hasard est précisément le nom qui désigne l'aptitude de la matière à s'organiser spontanément », Clément Rosset, *op. cit.,* p. 85.

d'intelligence artificielle : l'esprit « participe » de la matière, mais d'une façon qui invalide la conjecture de Dartmouth puisqu'il n'est pas possible d'en donner une description exhaustive. Les auteurs de la conjecture de Dartmouth ne peuvent pas postuler la primauté du hasard radical puisque cela consisterait à poser comme prémisse un « manque à penser », un vide, un reste, alors que le projet formulé à Dartmouth est justement de décrire l'intelligence *sans aucun reste*[592].

Admettre la prééminence du hasard sur la forme revient à affirmer qu'il n'y a pas de nécessité de la genèse des formes. Autrement dit la genèse s'engendre elle-même, ou encore la *formation* ne répond d'aucune *forme préconstituée*. Il est possible de décrire la pensée *dès qu'elle a pris forme* et d'expliquer, par exemple, la démarche de Lobatchevski. Mais la description manque toujours ce qui l'a fait advenir puisque ce n'est pas quelque chose d'extérieur, un « en plus », qui serait « intervenu » : cela s'organise de soi-même. Après coup, une forme spontanée est toujours descriptible comme un automatisme, mais sa spontanéité ne l'est pas. Il est possible de représenter les formes prises par la pensée, chaque façon qu'elle peut avoir de « s'organiser », mais pas son aptitude « à s'organiser spontanément ». Autrement dit, la créativité ne s'obtient pas en « ajoutant de la créativité » mais en laissant la place à du jeu, une « place » qui est une part mobile, sans règle de placement, sans « guidage » préalable, sans principe. La créativité ne répond à aucune règle, elle manifeste l'exception. Il ne peut y avoir de « règle de la créativité », puisque c'est l'exception qu'elle recherche, et cette exception, selon les mots de Godard, ne peut se « dire », elle peut seulement se manifester :

> […] tout dit la règle mais rien ne dit l'exception. Car il y a la règle et il y a l'exception. Il y a la culture qui est de la règle, et il y a l'exception, qui est de l'art. Tous disent la règle : cigarette, ordinateur, t-shirts, télévision, tourisme, guerre. Personne ne dit l'exception. Cela ne se dit pas, cela s'écrit – Flaubert, Dostoïevski – cela se compose – Gershwin, Mozart, cela se peint – Cézanne, Vermeer – cela se filme – Antonioni, Vigo – ou cela se vit et c'est alors l'art de vivre [...][593]

Si l'artiste doit faire preuve d'habileté, c'est pour suivre les directions données par le hasard, qui « seul préside aux destinées de la partie ». Sa « maîtrise » ne consiste pas à soumettre le hasard, (le « guider » par l'intuition, l'introduire épisodiquement dans une pensée « par ailleurs ordonnée »), mais au contraire à *mieux se soumettre au hasard*, c'est-à-dire faire preuve d'une

592 Pour rappel, la formulation de la conjecture de l'appel à projet de Dartmouth insiste sur l'exhaustivité : « *every aspect* of learning or *any other feature* of intelligence... », McCarthy et al., *op. cit.* Nous soulignons.
593 Jean-Luc Godard, *Je vous salue Sarajevo,* Toulouse, Éditions ECM Cinéma, 1993.

faculté à adapter sa production en fonction de bifurcations inattendues. Il lui faut une dextérité suffisante pour improviser sur le thème donné par les caprices du hasard.

Aussi en décrivant la créativité comme un jeu où « intervient » un hasard épisodique, les auteurs de la proposition de Dartmouth se trompent de jeu. Pour Clément Rosset, ce qui caractérise les jeux dits « de hasard » (les jeux provenant du château éponyme) est l'exclusion du recours à l'habileté et l'accent sur la passivité du joueur, « à qui était refusée toute possibilité d'*intervention* : 'hasard' seul présidait aux destinées de la partie[594]. » Cette passivité du joueur figure celle de la notion de hasard pur qui s'interdit tout recours extérieur « s'appelât-il chance, destin, providence ou fatalité[595] ». Un tel jeu provoque une terreur associée à « l'expérience de la perdition » : le joueur n'ayant plus aucun contrôle sur le cours des événements.

594 Clément Rosset, *op. cit.*, p. 77.
595 *Ibid.*

3.3. Paradoxes du contrôle

3.3.1. Le capitaine de son âme

À travers une étude sur les alcooliques, Bateson entreprend la critique de la notion de contrôle en s'attaquant au « concept occidental de *soi* », c'est-à-dire « une variante tout particulièrement catastrophique du dualisme cartésien : la division entre Esprit et Matière ou, en l'occurence, entre volonté consciente ou 'soi' (*self*) et le reste de la personnalité[596]. » L'alcoolique en état de sobriété « pense qu'il peut ou, du moins, doit être 'le capitaine de son âme[597]' » – une idée fausse dont il pâtit sans en avoir la responsabilité puisqu'il ne fait que souscrire à une prémisse partagée par ceux qui l'entourent. Aussi quand ceux-ci lui reprochent de ne pas savoir se contrôler, d'être « faible » ou « immature », ils ne font que renforcer « le style de sobriété qui le pousse à boire ». Lorsqu'il boit et renonce à « l'autocontrôle » et au combat contre lui-même, l'alcoolique « ne fait qu'apporter une correction (subjective) » de l'erreur, voire la « pathologie » de son style de sobriété – ou plutôt du style de sobriété de son groupe. « On pourrait aussi suggérer que l'alcoolique en état de sobriété est en quelque sorte plus sain d'esprit que ceux qui l'entourent et que cette situation lui est intolérable[598]. » Il souffre de « l'étrange épistémologie dualiste qui caractérise la civilisation occidentale[599] ».

Ainsi, « l'épistémologie de la maîtrise de soi, que ses amis infligent à l'alcoolique, est en elle-même monstrueuse. L'alcoolique a raison de la rejeter[600]. » Mais il ne peut la rejeter directement puisqu'elle structure sa propre façon de penser. Il peut, par contre, la prendre tellement au sérieux que son inconsistance finit par se manifester. L'alcoolique souscrit au mythe de l'autocontrôle et le met à l'épreuve (« juste un petit verre »), ce qui l'amène à chaque fois à perdre le contrôle.

---

596 Gregory Bateson, *Vers une écologie de l'esprit 1, op. cit.,* p. 270.
597 *Ibid,* p. 268.
598 *Ibid,* p. 267. Juste après avoir énoncé cette suggestion, Bateson revient en arrière : « J'ai entendu personnellement des alcooliques parler ainsi, mais je préfère ne pas prendre ici en ligne de compte une telle hypothèse. » Pourtant l'ensemble de l'article vise bien à montrer l'erreur du « style de sobriété » du monde occidental, c'est-à-dire le dualisme corps-esprit et l'idée d'autocontrôle, en proposant une notion alternative de « soi » où le « contrôle » s'effectue par circulation d'informations entre tous les éléments d'un système comprenant le corps et son environnement. Voir infra.
599 *Ibid,* p. 278.
600 *Ibid,* p. 285.

> En ce sens, la fierté de l'alcoolique est en quelque sorte ironique. C'est un effort résolu de vérifier la 'maîtrise de soi', avec un but ultérieur indicible, qui est de prouver en fin de compte qu'elle est inefficace et absurbe : « tout simplement ça ne marche pas[601] ».

Son expérience ne cesse de lui rapporter ce qui devient « un fait incontestable : mettre à l'épreuve la 'maîtrise de soi' conduit à nouveau à la boisson[602]. »

La maîtrise de soi est une prémisse que chaque soûlerie vient démentir en prouvant à l'alcoolique qu'il n'est pas le « capitaine de son âme » puisque l'alcool décide à sa place. Mais tant que l'alcoolique souscrit à la notion de maîtrise de soi, il trouve un moyen d'interpréter les événements de façon à éviter ce constat : puisqu'il est indiscutablement le « capitaine de son âme » ce sont des circonstances extérieures qui l'ont amené à perdre le contrôle. La prémisse restant constitutive de son « épistémologie », l'alcoolique ne peut qu'interpréter les événements d'une façon qui la préserve au lieu de la remettre en question. Chaque mise à l'épreuve de sa capacité de maîtrise répète l'expérience de l'échec mais sera interprétée comme le fait d'une cause extérieure, ce qui rétrécit d'autant sa « zone de maîtrise » imaginaire : « la 'fierté' alcoolique rend le concept de soi de plus en plus étroit, en plaçant à l'extérieur de son champ une grande partie de ce qui se passe[603] » et en structurant la relation entre l'alcoolique et cet « extérieur » comme une rivalité[604]. *In fine*, l'alcoolique en arrive à penser que tout est contre lui :

> Le principe de la fierté-dans-le-risque est en fin de compte plutôt suicidaire. Libre à vous de vouloir vérifier une fois si l'univers est de votre côté ; mais remettre ça sans cesse, tenter une concertation croissante des preuves en ce sens, c'est se laisser aller à un projet qui, mené à son bout, ne peut prouver qu'une seule chose : à savoir que l'univers vous hait[605].

Comment s'en sortir si l'alcoolique n'est pas le « capitaine de son âme » et si sa situation empire quand il prétend reprendre le contrôle ?

Citant l'organisation des Alcooliques Anonymes, Bateson rapporte que le changement ne peut avoir lieu que si l'alcoolique « touche le fond » afin que « le mythe de la maîtrise de

601 *Ibid,* p. 285.
602 *Ibid.*
603 *Ibid,* p. 280.
604 Bateson détaille ce point en s'appuyant sur la distinction entre relation symétrique et relation complémentaire. L'alcoolique entre dans une relation de rivalité symétrique avec la bouteille et plus généralement avec son environnement.
605 *Ibid.*

soi du sujet [soit] démoli[606] ». En d'autres termes, ce n'est pas l'alcoolique qui choisit. L'organisation des Alcooliques Anonymes « considère qu'il y a peu de chances de venir vraiment en aide à un alcoolique qui n'a pas encore touché le fond ». Seuls les événements peuvent l'amener jusqu'au point où il est « vaincu par la bouteille et en [est] conscient », étape qui peut à la fois « convaincre l'alcoolique qu'un changement est nécessaire » et « *est* elle-même la première étape de ce changement[607]. » Il n'y a pas de règle générale quant à cet événement déclencheur, il dépend de chaque individu et des circonstances – il est *contingent* :

> attaque de delirium tremens, un laps de temps pendant une soûlerie dont il n'a aucun souvenir, le rejet de la part de sa femme ou la perte de son travail, un diagnostic sans espoir, etc. – autant d'événements qui peuvent avoir l'effet requis. « AA » [Alcooliques Anonymes] affirme que le « fond » est différent pour chaque individu et que certains seront morts avant d'avoir atteint au leur[608].

Si l'alcoolique a touché le fond, l'organisation des Alcooliques Anonymes peut essayer de lui faire admettre qu'il est sans « défense devant l'alcool[609] », c'est-à-dire de lui faire renoncer au principe de la maîtrise de soi. Mais la contingence du « toucher le fond » est aussi contingence des effets de l'événement : « Cette étape est communément considérée comme une 'reddition', et souvent les alcooliques sont incapables de la suivre jusqu'au bout ou, sinon, ils ne la réalisent que temporairement, pendant la période de remords qui fait suite à une débauche[610]. » Le risque est grand qu'« après une période plus ou moins longue de sobriété il[s] tente[nt] à nouveau d'utiliser l'autocontrôle afin de combattre la 'tentation'[611]. »

Avec le revirement de l'alcoolique, Bateson entend donner un exemple de *changement d'épistémologie :*

> D'un point de vue philosophique, cette première étape ne constitue nullement une reddition, mais un changement d'épistémologie, un changement d'appréhender la personnalité dans son propre monde. C'est ce changement qui s'effectue d'une épistémologie incorrecte vers une autre plus correcte[612].

606 *Ibid*, p. 269.
607 *Ibid.*
608 *Ibid*, p. 288.
609 *Ibid*, p. 269.
610 *Ibid.*
611 *Ibid.*
612 *Ibid*, p. 270.

Le changement s'effectue involontairement puisqu'il s'agit justement d'abandonner le principe de la maîtrise de soi. Pour discréditer le « concept occidental de 'soi' », Bateson montre que nous agissons *selon* des principes[613] ; que ces principes changent ; et que ce changement ne peut être volontaire (puisque les principes prédéterminent toute action « volontaire »), il ne peut se faire qu'à la faveur d'événements contingents. Avec facétie, il illustre ce changement de principe involontaire par un cas d'abandon salutaire du principe de volonté, ici chez l'alcoolique.

3.3.2. Décapiter l'esprit

Avec la notion de « fierté de soi » de l'alcoolique, Bateson met en lumière un cas où *plus la maîtrise de soi est mise à l'épreuve* – plus nous cherchons à maîtriser la maîtrise de soi – et moins nous nous maîtrisons. En faisant de l'alcoolique celui qui souffre du « concept occidental de *soi* », Bateson laisse entendre que le paradoxe vaut également pour les non alcooliques.

Si, à l'instar de l'alcoolique, je mets ma « maîtrise de soi » à l'épreuve en me plaçant dans une situation où celle-ci est convoquée, je risque de ne rien trouver. C'est au moment les plus cruciaux, ceux où mon effort pour le contrôler est le plus grand (examen, prise de parole en public, entretien d'embauche, rendez-vous amoureux…) que mon corps me trahit : il transpire abondamment et répand une odeur infame, ou alors j'ai des gaz, je renverse une tasse de thé, et mille et une autres façons de « se prendre les pieds dans le tapis ». Si je tourne mon regard vers ce qui est censé contrôler mon corps – mes pensées, le siège de ma volonté et de mon libre-arbitre – je n'ai pas de meilleur constat. Dans ces moments de mise à l'épreuve, je rate la station de métro (distraction), je me trompe de nom pour m'adresser à un examinateur (lapsus), je ne donne pas la bonne réponse alors que je la connais, je me rends compte après-coup de ce qu'il aurait fallu dire, etc. Si je redoute autant ces moments d'examen, de cérémonie, c'est que je sais que je ne ferai probablement pas l'expérience d'un corps et d'une pensée qui « obéissent » à mes injonctions. Au contraire, je vais appréhender les épreuves avec anxiété par anticipation des surprises qu'ils susciteront.

613 « tous les êtres humains (et, en fait, tous les mammifères) sont guidés par des principes hautement abstraits, dont ils sont presque entièrement inconscients, ignorant que le principe qui gouverne leurs perceptions et actions est d'ordre philosophique. » *Ibid,* p. 278.

Je ne peux pas refuser à une pensée d'avoir lieu. Imaginons que je me trouve dans un musée face à une sculpture dans un matériau que je n'ai jamais vu, et qui semble très doux. Je peux me retenir de toucher cette sculpture, je peux censurer un geste ou un mot, mais pas la pensée qui me vient de toucher cette sculpture. Une fois que la pensée a eu lieu, je pourrais chercher à l'interrompre, notamment en dirigeant mon esprit vers autre chose, et en me concentrant pour ne pas me laisser distraire. Mais cela a toujours lieu après. Je ne peux pas m'interdire de penser à toucher l'œuvre *avant* d'y avoir pensé. Au moment où je me l'interdit, je la pense. Je ne peux pas ne pas penser à ce qu'il est interdit de penser pour m'interdire de le penser. De la même façon, si je visite le musée avec un enfant, je peux lui interdire de toucher la sculpture, je ne peux lui interdire d'y penser. Au moment où je formule l'interdiction – « je t'interdit de penser à toucher cette sculpture » - voilà qu'il y pense, quand bien même son esprit aurait été occupé à tout autre chose.

Avant de signifier « direction » ou « maîtrise », le mot contrôle désigne la « vérification », et plus précisément le fait, pour une vérification, de *doubler* l'enregistrement des faits : un contrôle est ainsi un « registre tenu en double, l'un servant à vérifier l'autre appelé rôle[614]. » Le mot a perdu cet usage, mais il faut toujours, pour qu'il y ait contrôle, qu'il y ait un enregistrement et un dédoublement : ce qui contrôle doit se distinguer de ce qui est contrôlé pour le contrôler. Dans « l'autocontrole » je me mets en écart à moi-même. Mais alors que je peux étendre mon bras et contrôler simultanément que je l'étend de la bonne façon, je le contrôle avec la pensée, via la médiation de la vue. Mais je ne peux pas faire cela vis-à-vis de mes pensées. Si je décide de me concentrer sur un problème et de ne pas me laisser distraire, cette pensée est déjà une distraction : au moment où j'ai cette pensée, je ne suis pas entièrement concentré sur le problème. Si je veux me reconcentrer parce que j'ai été distrait, c'est que j'ai déjà été distrait.

La pensée ne semble pas présenter de médiation entre le désir et l'acte : si je veux penser à quelque chose, je l'ai déjà pensé. Alors que si je veux toucher la sculpture, il me faudra m'en approcher, lever le bras, peut-être déjouer la surveillance du gardien. Il y aura un délai, et peut-être une résistance des choses, qui viendront s'interposer, servir de médiation entre le désir et l'action, et donner une durée au désir. J'éprouve mon désir pendant ce délai, et je l'éprouve d'autant plus que les choses y résistent – ce qui peut même devenir pénible. En ce qui concerne la pensée, l'absence de médiation ne laisse pas à la volonté le temps de s'éprouver. Si bien qu'on

614 « contrôle », *Centre national de ressources textuelles et lexicales*, https://www.cnrtl.fr/definition/contr%C3%B4le page consultée le 20 janvier 2019.

peut douter que celle-ci intervienne vraiment : comme si la pensée était une « passivité », plutôt qu'une activité.

Plus généralement, pour peu que j'observe mes pensées j'y verrai beaucoup de vieilles rengaines, répétitions d'événements passés ou anticipations d'événement futurs, ainsi que des scénarios imaginaires très élaborés correspondant à une foule de désirs et de colères que j'aurais du mal à assumer s'il fallait les rendre publics, mais assez peu de délibérations rationnelles. Je trouve dans ce parlement intérieur plus de lobbyistes (les lobbies de la nourriture, de la vengeance, de la sieste…), que de dirigeants éclairés, à tel point que je peux me demander comment est-ce qu'une république si mal fagotée a pu malgré tout garder une direction conforme à la norme et à mon intérêt à long terme. Je constate en moi une variété d'élans intempestifs qui se sont constitués au fil de mon expérience en un système de préférences plus ou moins cohérent – sans que je puisse en décider.

Ainsi, à mesure que j'observe mon environnement, mon comportement, mes pensées de façon à délimiter *sur quelle zone* s'exerce ma maîtrise, je verrai une réduction progressive du « soi » que je suis censé maîtriser : jusqu'à mon « intériorité », les pensées de mon « for intérieur », résistent à une description en termes de contrôle, et cela d'autant plus aux moments où j'en ai besoin, c'est-à-dire quand le contrôle est mis à l'épreuve.

Si c'est au moment où il faut faire preuve d'autocontrôle, où le regard se porte sur lui, qu'on ne le trouve pas, cela veut-il dire que pour être assuré de ne pas perdre le contrôle, il faudrait ne pas chercher à contrôler ? Quelle est cette administration folle qui n'obéit *qu'à condition qu'on ne lui donne aucun ordre* – qui aurait pour règle de *moins obéir* à mesure que le roi commande ? Et à quoi obéit-elle alors ? Quel est ce pouvoir étrange dont l'efficacité est exactement inverse aux efforts déployés pour l'exercer ? Et qu'est-ce alors que le « soi », ce roi déficient dont le pouvoir ne peut s'exercer qu'involontairement ?

Ces moments où le contrôle est absent pourraient être interprétés comme de simple « pertes de contrôle ». C'est la panique qui fait « perdre » un contrôle qui, le reste du temps, est bien présent. En d'autres termes, les moments d'examen ou de rendez-vous amoureux seraient une *exception* à la règle qui veut que, la plupart du temps, nous nous contrôlons. Et cette exception, en montrant ce qui se passe quand le contrôle défaille, serait la meilleure façon de manifester la réalité de l'autocontrôle puisque le reste du temps nous ne nous prenons pas sans cesse les pieds dans le tapis. Cette suspension de l'autocontrôle serait la meilleure façon de confirmer son existence.

Mais est-il pertinent de penser que je fais preuve d'autocontrôle quand, au quotidien, je ne me prends pas les pieds dans le tapis ? Ce matin, je n'ai ni renversé ma tasse de thé ni fait

de lapsus en parlant à mon chef. Mais je n'ai fait aucun effort pour éviter de laisser passer un lapsus ou de renverser ma tasse. A vrai dire, je n'ai même pas pensé à ces éventualités. En quoi est-ce un autocontrôle si celui-ci est involontaire ?

Nous avons vu qu'au quatorzième siècle, le « role » désigne un « rouleau » ou registre que le « contre role » vient recopier pour vérification. Si aujourd'hui le terme a pris un sens plus vaste de vérification et de filtrage (« contrôle des sacs à l'entrée »), il a pour origine une mise en registre, un enregistrement du « rôle » qui vise à s'assurer que celui-ci correspond bien à ce qui est attendu. Pour qu'il y ait auto-contrôle il doit y avoir une sorte de dédoublement : une partie de moi observe mon comportement. Elle enregistre le rôle. Mais si je suis mis à l'épreuve et que j'ai un doute sur la situation, je vais aussi contrôler le contrôle : cette partie de moi qui évalue le comportement, est-ce qu'elle l'évalue de la bonne façon ? Ai-je raison de penser que « je suis trop tendu » pour ce rendez-vous ? Il peut y avoir des dédoublements successifs me permettant de contrôler les contrôles, mais à aucun moment il n'y aura d'autocontrôle complet puisque la partie qui contrôle le reste devrait aussi se contrôler elle-même pour contrôler la totalité du soi. S'il y a dédoublement – pour effectuer le contrôle –, alors cela implique une partie contrôlante qui n'est pas contrôlée. Il ne peut donc y avoir que des *contrôles partiels* qui s'appliquent à des processus isolés et non autocontrôle s'appliquant au « soi ».

A cela s'ajoute que le contrôle a lieu de façon *décalée* : il est anticipé, lorsque je m'interdis une action, ou retardé, lorsque j'observe mon rôle et je l'évalue après coup. Il faut toute l'élaboration d'un double de la chose contrôlée, du « rôle », qu'on évalue ensuite par rapport à un idéal (« ce qui est attendu »), mais il n'y a pas de *simultanéité* entre le contrôle et le registre contrôlé. Je ne peux pas contrôler ma pensée en même temps que je la pense. Je ne peux pas avoir deux flux de pensée à la fois, l'un contrôlant l'autre. Si je décide de me concentrer sur un problème et de ne pas me laisser distraire, cette pensée est déjà une distraction. Si je veux me reconcentrer parce que j'ai été distrait, c'est que j'ai déjà été distrait.

De ce point de vue, le moment de panique où je réalise le décalage de mes gestes, paroles et pensées par rapport à ce qui en est attendu n'est pas la perte momentanée d'un contrôle que j'exerce le reste du temps, ce n'est pas une exception au régime habituel du contrôle, mais la révélation de l'impossible simultanéité du contrôle et de l'action. Bateson compare l'alcoolique qui entre dans un bar à quelqu'un qui essayerait de freiner sur une route verglassée. « On peut comparer ce test à une autre situation : celle d'un conducteur auquel on demanderait de freiner sec sur une route glissante : il découvrira sans tarder que son contrôle sur sa voiture est

limité[615]. » De la même façon, si nous ne mettons pas en doute l'autocontrôle au quotidien, c'est que nous ne donnons pas de coup de frein, nous ne cherchons pas à reprendre le contrôle du véhicule. Aussi l'objet de notre panique est-il plus général : l'autocontrôle est absent au moment de l'épreuve, *révélant qu'il n'a jamais été présent*.

Pour autant, il existe de nombreux exemples de manifestations que nous prenons pour de l'autocontrôle, comme la performance d'un danseur ou d'une pianiste. Mais c'est moins l'effet d'une volonté précise dans l'instant que d'une longue pratique d'incorporation d'automatismes. Celui qui maîtrise un geste est justement celui qui peut se passer de contrôle. Il a confiance dans l'adéquation entre son geste et la situation et n'a pas besoin d'anticiper ou de revenir sur son geste. Ici, la « maîtrise » est incorporation d'automatismes et non maîtrise « de soi ». À cela s'ajoute que ce déploiement de « maîtrise » est rendu difficile pour peu que l'on perturbe le contexte, comme si ce dernier participait à l'effectuation de ces automatismes. Pour rendre compte du « soi » que semble maîtriser le pianiste, il faudrait y inclure le piano.

Pour Bateson, il ne peut y avoir de « maîtrise de soi » dans la mesure où la notion de « soi » « n'est qu'une petite partie d'un système beaucoup plus vaste *d'essais-et-d'erreurs,* à travers lequel s'opèrent la pensée, l'action et la décision[616]. » Autrement dit, « le 'soi' est une fausse réification d'une partie mal délimitée de cet ensemble beaucoup plus vaste de processus entrelacés[617]. » Lorsqu'un humain abat un arbre, nous simplifions le processus à l'extrême en le réduisant à un esprit dans un humain qui agit sur l'arbre :

> le parler courant exprime l'*esprit (mind)* à l'aide du pronom personnel, ce qui aboutit à un mélange de mentalisme et de physicalisme qui renferme l'esprit dans l'homme et réifie l'arbre. Finalement l'esprit se trouve réifié lui-même car, étant donné que le 'soi' agit sur la hache qui agit sur l'arbre, le 'soi' lui-même doit être une 'chose'[618].

Mais *ce qui abat l'arbre* est un système plus vaste que le soi réifié. Pour mieux le décrire, il faudrait considérer le circuit arbre-yeux-cerveau-muscle-cognée-coups-arbre et les « conversions de différences » qui se transmettent le long du circuit et constituent un processus autocorrecteur : « chaque coup de cognée sera modifié (ou corrigé) en fonction de la forme de l'entaille laissée sur le tronc par le coup précédent[619]. »

615 Gregory Bateson, *op. cit.,* p. 289.
616 *Ibid,* p. 290.
617 *Ibid.*
618 *Ibid,* p. 275.
619 *Ibid,* p. 274.

A partir du moment où l'on cherche à délimiter précisément le soi, l'inconsistance de la notion est manifeste. Pour un aveugle qui ne peut se déplacer sans sa canne, tout se passe comme si cette dernière faisait partie du « soi » : « la canne est tout simplement une voie, au long de laquelle sont transmises les différences transformées, de sorte que couper cette voie c'est supprimer une partie du circuit systémique qui détermine la possibilité de locomotion de l'aveugle[620]. » Le point de vue correct sur l'action est donc celle qui prend en compte le circuit dans son ensemble et non le « soi » opposé au corps et à l'arbre. À la « réification du soi » et au dualisme, Bateson oppose une définition de l'esprit comme système sensible aux différences, une « idée élémentaire » étant pour lui une « différence qui se déplace et subit des modifications successives dans un circuit[621] ».

Une telle définition fait s'effondrer la notion de « maîtrise de soi » puisqu'« aucune partie de ce système intérieurement (inter) actif *ne peut exercer un contrôle unilatéral* sur le reste ou sur toute autre partie du système[622]. » Bateson illustre ce point en décrivant comment le régulateur d'une machine à vapeur n'exerce pas de « contrôle unilatéral » mais est un « organe sensible » qui traduit l'écart entre une vitesse préférable et une vitesse effective en modulant l'arrivée de combustible et le freinage. Il est incorrect de penser que le régulateur détermine le système puisque « le comportement du régulateur est déterminé par le comportement des autres parties du système et indirectement par son propre comportement à un moment antérieur[623]. » *Ce qui détermine l'action* n'a pas les mêmes limites que ce que l'on appelle généralement le « soi », et plus généralement, cela n'a pas de sens de considérer une sous-partie du système comme déterminant les autres.

Pour Bateson, la cybernétique est aussi ironique que la « fierté de l'alcoolique ». C'est une « science du contrôle » et du gouvernement (*kubernetes*) qui en vient à montrer qu'il n'y a pas de gouverneur. Les exemples mentionnés de mise à l'épreuve du « gouvernement de soi » montrent qu'il peut y avoir contrôle (au sens où j'enregistre mon rôle et le compare à un idéal) mais pas de maîtrise par le soi (puisque l'adaptation de mes gestes ne vient pas de moi). Au contraire, il peut arriver que le sentiment de maîtrise diminue à mesure que j'instaure un

620 *Ibid,* p. 275.

621 « […] tout système fondé d'événements et d'objets qui dispose d'une complexité de circuits causaux et d'une énergie relationnelle adéquate, présente à coup sûr des caractéristiques 'mentales'. Il *compare,* c'est-à-dire qu'il est sensible et qu'il répond aux *différences* (ce qui s'ajoute au fait qu'il est affecté par les causes physiques ordinaires telles que l'impulsion et la force). Un tel système 'traitera l'information' et sera inévitablement autocorrecteur, soit dans le sens d'un optimum homéostatique, soit dans celui de la maximisation de certaines variables. Une unité d'information peut se définir comme une différence qui produit une autre différence. Une telle différence qui se déplace et subit des modifications successives dans un circuit constitue une idée élémentaire. » Gregory Bateson, *op. cit.,* p. 272.

622 *Ibid,* p. 272. Nous soulignons.

623 *Ibid,* p. 273.

contrôle, en particulier si je suis hors du contexte où mes automatismes ont l'habitude de se déployer en coopération avec les objets familiers. Il n'y a pas de « perte de contrôle » et refus du corps ou des pensées d'obéir aux ordres du « soi » mais un « manque d'habitude ». Je n'agis pas sur commande du « soi » mais via des dispositions acquises constituées par adaptation aux différents contextes où je me trouve.

3.3.3. Déterminisme et contingence des automatismes (l'habitus)

Ce n'est donc pas le « soi », un hypothétique « gouverneur » logé dans son for intérieur, qui fait la spécificité d'une personne, mais la collection d'automatismes acquis au cours de son existence lui permettant d'évoluer avec aisance dans différents contextes. La plupart de mes comportements viennent d'usages inculqués par ma famille et ma fréquentation de différents groupes. De la même façon, la plupart de mes décisions se sont jouées au sein d'un cadre socialement prédéterminé. J'ai peut-être « choisi » mes études, mais je n'ai pas choisi *le fait de faire des études,* puisqu'à aucun moment je n'ai interrogé la pertinence des études : ce genre de croyances « vont de soi », c'est-à-dire qu'elles viennent avec l'ensemble d'usages socialement incorporés. Bateson dresse un portrait des humains comme agissant selon un complexe de principes qu'il appelle leur « épistémologie » et qu'il définit comme les « règles dont se sert l'individu pour 'interpréter' son expérience », ou encore « un ensemble d'hypothèses ou de prémisses habituelles, implicites dans la relation entre l'homme et son environnement » qui « gouvernent l'adaptation (ou la non-adaptation) à l'environnement humain et physique[624] ». L'ensemble de « prémisses » permet d'attribuer une valeur à chaque expérience et d'orienter l'action. C'est en ce sens qu'il gouverne ou dirige celle-ci. Notre comportement se déduirait de cet ensemble de « prémisses » tout comme les théorèmes d'un système mathématique se déduisent d'un corpus d'axiomes, à la différence que notre « épistémologie » doit composer avec le changement, elle est en prise dans le temps. Ainsi l'« épistémologie » serait la façon dont nous *réglons* notre comportement, au sens où nous assimilons des règles mais aussi au

624 « c'est un ensemble d'hypothèses ou de prémisses habituelles, implicites dans la relation entre l'homme et son environnement, et que ces prémisses peuvent être vraies ou fausses. J'utiliserai donc ici le seul terme d' 'épistémologie' pour désigner les deux aspects des prémisses qui gouvernent l'adaptation (ou la non-adaptation) à l'environnement humain et physique. Pour reprendre l'expression de George Kelly, ce sont là des règles dont se sert l'individu pour 'interpréter' son expérience. » Gregory Bateson, *op. cit.,* p. 271.

sens où celles-ci sont modulées au fil de l'expérience : il y a une adaptation, un *réglage*. Bien que nous fondions nos actions et réflexions sur elles, ces prémisses sont implicites et ne sont pas forcément connues ou reconnues. Les prémisses ont été assimilées sans vérification de leur cohérence entre elles et peuvent donc être contradictoires[625], comme le montre Bateson en détaillant celles qui régissent le style de vie de l'alcoolique.

La notion d'« épistémologie » que propose Bateson peut être complétée par celle d'« habitus », élaborée par Bourdieu. Celle-ci désigne également un système de règles : les « principes générateurs des pratiques[626] » et le « répertoire de règles[627] » qui gouvernent l'action. Au terme de « règles », Bourdieu privilégie cependant le terme, plus souple, de « disposition », qui a l'avantage de combiner l'idée d'ordre ou arrangement, avec celui de tendance ou préférence, et d'inscrire l'habitus dans la pratique. Bourdieu insiste sur l'absence de distinction entre théorie et pratique[628] : il ne s'agit pas d'un système de règles logé « dans notre tête » qui régirait nos actions mais d'un ensemble d'inclinations qui se font et se défont par l'expérience – « un système de dispositions durables et transposables[629] » littéralement *incorporées* :

> En outre, tous les principes de choix sont incorporés, devenus postures, dispositions du corps : les valeurs sont des gestes, des manières de se tenir debout, de marcher, de parler. La force de l'ethos, c'est que c'est une morale devenue hexis, geste, posture[630].

Ce n'est donc pas une morale abstraite qui pourrait être traduite dans un jeu de règles explicites mais un ensemble d'attitudes, de manières d'êtres, de réflexes, d'automatismes physiques ou mentaux. La plupart de ces automatismes sont transmis – l'habitus prend sa source dans un milieu social donné et permet d'avoir un « sens pratique » de ce milieu. Il fournit une faculté d'adaptation immédiate à ce qui s'y effectue.

---

625 Bateson précise que « ces prémisses peuvent être vraies ou fausses ». *Ibid.*

626 Pierre Bourdieu, *Esquisse d'une théorie de la pratique,* Genève, Librairie Droz, 1972, p. 163.

627 *Ibid*, p. 159.

628 « La notion d'habitus englobe la notion d'ethos, c'est pourquoi j'emploie de moins en moins cette notion. Les principes pratiques de classement qui sont constitutifs de l'habitus sont *indissociablement* logiques et axiologiques, théoriques et pratiques (dès que nous disons blanc ou noir, nous disons bien ou mal). La logique pratique étant tournée vers la pratique, elle engage inévitablement des valeurs. C'est pourquoi j'ai abandonné la distinction à laquelle j'ai dû recourir une fois ou deux, entre eidos comme système de schèmes logiques et ethos comme système des schèmes pratiques, axiologiques (et cela d'autant plus qu'en compartimentant l'habitus en dimensions, ethos, eidos, hexis, on risque de renforcer la vision réaliste qui porte à penser en termes d'instances séparées). » Pierre Bourdieu, « Le marché linguistique », in *Questions de sociologie*, Paris, Éditions de Minuit, 1992, p. 133-134.

629 Pierre Bourdieu, *Esquisse d'une théorie de la pratique, op. cit.,* p. 178.

630 Pierre Bourdieu, *Questions de sociologie, op. cit.,* p. 134.

Les actions ne sont pas déduites d'un système de règles logé à part dont on pourrait évaluer la cohérence. C'est la situation qui requiert l'automatisme correspondant, la diversité des situations supposant une diversité d'automatismes qui, si on s'attarde à les justifier, peuvent renvoyer à des principes disparates. Contrairement aux notions de règles et de prémisses, l'idée d'un ensemble de dispositions incorporées mobilisées par le contexte laisse plus de latitude à la cohabitation de dispositions disparates. Au fil de son histoire, chaque individu est plus ou moins amené à fréquenter différents « mondes » ou « milieux » et à incorporer différentes dispositions hétérogènes, voire contradictoires. « La diversité des expériences sociales (ascension sociale ou déclassement, hétérogamie, etc.) peut ainsi générer des habitus individuels clivés ou dissonants[631] ». L'habitus tolère une dose de contradiction.

Cet ensemble de dispositions s'élabore et évolue sans que nous ayons la main dessus. Je ne peux pas en décider puisque c'est justement ce qui me fait décider, ce par quoi je prends mes décisions. Je n'ai pas de référentiel de valeurs qui me permettrait de faire évoluer mon référentiel de valeurs. Ces dispositions semblent aussi inscrites que les appétits et dégoûts que j'ai incorporés. Je ne peux pas décider du jour au lendemain de devenir de droite ou d'aimer le fromage. Et pourtant l'*habitus* évolue. Chaque événement infléchit et module la relation au monde, avec des évolutions imperceptibles et de grandes ruptures. On s'embourgeoise ou on se radicalise, ici l'appétit s'émousse et là le goût s'affine. Au fil des rencontres, on acquiert et on perd des manières d'être.

Dans la mesure où l'habitus est modelé par l'expérience et dépend du milieu de cette expérience, il est le relai du déterminisme social. Nous ne pouvons pas le modifier à notre guise puisque « notre guise » en est le produit. Le mieux que nous puissions faire est de prendre du recul par rapport à cet ensemble de dispositions. La reflexivité permet d'introduire du jeu, une suspension partielle des principes leur permettant d'évoluer, tout comme le retour sur les axiomes opéré par Lobatchevski a permis de faire évoluer la géométrie.

La genèse de cette « épistémologie » ou *habitus* se fait par une éducation au sens large. Ce ne sont pas seulement les règles imposées par la famille ou l'école, mais les leçons inculquées par l'ensemble des expériences vécues. Il y a une éducation – ou *paideia*[632] – qui est aussi une *paidia*, puisqu'il y a passage d'un jeu de règles (ou ensemble de dispositions) à un autre. La sociologie met l'accent sur le déterminisme social qui permet de rendre compte d'un

631 Anne-Catherine Wagner, « habitus » *in* Serge Paugam (dir.), *Les 100 mots de la sociologie*, Paris, Presses universitaires de France, coll. « Que Sais-Je ? », 2018.

632 Le terme de παιδεία désigne l'éducation dans le monde Grec Antique. Werner Jaeger, *Paideia : la formation de l'homme grec,* Paris, Gallimard, 1988.

habitus à un moment donné, mais *l'évolution* de celui-ci se fait au gré des événèments et des rencontres, petits ou grands, heureux ou malheureux. Nous retrouvons ici la distinction entre le *ludus* ou jeu de règles fixes, qui peut s'appliquer à l'habitus à un instant donné, et la *paidia* ou fantaisie sans règles, qui s'applique à son évolution au gré des rencontres. Selon l'origine sociale d'un individu, on peut lui attribuer une probabilité plus ou moins grande de vivre tel ou tel événement (passer son bac, prendre un crédit, devenir ouvrier ou cadre…), mais la suite d'événements qu'il vivra effectivement demeure contingente. À l'instar de la thermodynamique des gaz, la trajectoire d'un ensemble d'individus peut suivre des lois statistiques sans pour autant que chaque trajectoire particulière soit déterminée[633]. Une partie de l'habitus participe de dispositions communes : celles de l'époque, du lieu et du milieu social. Mais chaque habitus individuel étant modelé selon sa propre suite d'événements contingents, il est singulier.

La notion d'habitus permet de préciser ce que Bateson appelle l'« épistémologie ». Les principes qui président à l'action et au choix sont moins des « prémisses » que des « dispositions » inculquées par l'éducation et l'habitude. Ils ne sont pas « dans la tête » mais sont incorporés. « On pourrait, déformant le mot de Proust, dire que les jambes, les bras, sont pleins d'impératifs engourdis[634]. » On comprend mieux comment « le mental » (et donc ce à quoi on attribue l'action) peut être distribué dans le corps et ce qui l'entoure :

> Ainsi, dans aucun système qui fait preuve de caractéristiques « mentales », n'est donc possible qu'une de ses parties exerce un contrôle unilatéral sur l'ensemble. Autrement dit : *les caractéristiques « mentales » du système sont immanentes, non à quelques parties, mais au système entier*[635].

En tant qu'ensemble d'automatismes acquis par l'expérience et permettant de répondre aux situations *sans délibération,* l'habitus n'est pas sans ressembler aux réseaux de neurones. Il est tentant de comparer AlphaGo et l'habitus dont Bourdieu précise qu'« il est au principe de ces enchaînements de *coups* qui sont objectivement organisés comme des stratégies sans être le produit d'une véritable intention stratégique[636] ». Mais la comparaison ne tient que pour un instant et une situation donnée. Il y a *ludus* lorsqu'un milieu et une situation identifiés impose

633 C'est d'ailleurs en s'inspirant de la sociologie naissante que Maxwell défend l'idée qu'il n'est pas nécessaire d'étudier les trajectoires individuelles des particules pour connaître leur comportement global. Avec Bolzmann, puis Gibbs, ils auront le plus grand mal à faire reconnaître la légitimité de cet écart par rapport au déterminisme qui avait caractérisé la physique jusqu'alors. Voir Olivier Rey, *Quand le monde s'est fait nombre,* Paris, Stock, 2017, chapitre sept.

634 Pierre Bourdieu, *Le Sens pratique*, Paris, Éditions de Minuit, 1980, p. 117.

635 Gregory Bateson, *op. cit.*, p. 273.

636 Pierre Bourdieu, *op. cit.*, p. 103-104.

une cohérence aux actions. On peut « jouer à la marchande » ou « au docteur ». La comparaison ne tient plus dès qu'on prend en compte le fait que l'habitus peut intégrer des automatismes contradictoires (habitus clivé) et le fait qu'il évolue. Il a une histoire qui suit un déroulement non programmé.

3.3.4. Qu'est-ce que « changer » ? S'oublier dans la révolution

Pour François Roustang, c'est une forme de narcissisme qui nous pousse à croire que le soi est une entité fixe qui contrôle le corps. Nous aimons nous comparer aux machines car elles ont un plan déterminé, elles se laissent « analyser et synthétis[er] », « à loisir[637] ». C'est un miroir dans lequel se contempler. Mais la comparaison ne provoque que de la souffrance. Le soi n'ayant jamais la consistance et la cohérence des machines auxquelles il est comparé, nous voilà persuadés que nous dysfonctionnons. Nous n'avons de cesse de réparer le soi, de lui trouver enfin une cohérence. C'est une quête vaine, qui suscite des souffrances inutiles. En l'entretenant, la psychanalyse risque de jouer le rôle de *pharmakon*. En assimilant le soi à une machine, on se prive de comprendre comment il peut changer, et on s'éloigne d'autant de ce que l'on attend pourtant de la thérapie – le changement.

Les mutations du soi – changement de paradigme, d'épistémologie ou d'habitus – ne pouvant être comprises qu'après coup, la thérapie se doit au contraire de permettre un relatif *oubli de soi*. Chercher, avant qu'il n'ait eu lieu, à comprendre le changement qui s'opère, risque de l'entraver, de le compliquer, de le rendre plus douloureux encore. Le changement se fait malgré nous, ou en tout cas malgré la partie de nous qui cherche à comprendre, voire à contrôler, ce qui a lieu. Le changement se fait en partie à notre insu, ce qui suscite de l'angoisse. Pour changer, il faut donc « s'oublier », comme on dit d'un grabataire qu'il s'est oublié dans son lit, s'offrir un laisser-passer vers « le soi d'après », qui est inimaginable pour « le moi d'aujourd'hui ».

Ce qui est à oublier, c'est ce qui en nous ratiocine, « celui-là qui parle, réfléchit et décide » et qui « ne cesse d'osciller entre la crispation et l'incertitude concernant son individualité[638] ». Celui-là doit en rester à un rôle de spectacteur, de facilitateur, au risque de se

637 François Roustang, *Influence*, Paris, Éditions de Minuit, 1991, p. 15.
638 *Ibid*, p. 12.

mettre en travers du cours des événements et d'en souffrir. Nous retrouvons l'idée évoquée lors de nos réflexions sur l'invention et la découverte : lorsque la conscience risque d'entraver l'apparition de quelque chose de nouveau, une suspension temporaire, un bref *aveuglement,* peut permettre de faciliter le processus. Chaque changement est comme une révolution que le sujet conscient doit laisser advenir bien que celle-ci le remette en cause. Une fois le changement advenu, il ne pourra qu'abandonner le paradigme qui donnait du sens à ce qui lui arrivait, abandonner le récit de soi par lequel il entretenait la consistance et la cohérence de son image. Chaque changement réalise un événement impossible du point de vue du moi spectateur – son « paradigme » est pris en défaut –, et c'est en même temps un retour à la simplicité du présent depuis la perception du corps. Il y a révolution dans les deux sens du terme : changement radical (rupture avec le passé) et retour au même (la banalité du présent). Pour que cette révolution puisse avoir lieu, il faut abandonner la notion de maîtrise, marcher comme Saint-Denis avec la tête entre les mains. Plutôt qu'une *action,* le changement implique donc un *retrait,* retrait qui permet que les événements prennent un nouveau cours, comme la suspension d'un axiome par Lobatchevski a permis à la géométrie de prendre un nouveau cours.

3.3.5. Le sujet sans la maîtrise : de la quête de l'absolu à la diplomatie des liens

L'objectif de François Roustang est d'élaborer une critique de la théorie freudienne, en particulier de la conception du sujet comme machine close. S'appuyant sur la notion d'influence, et soulignant l'ambiguité de Freud vis-à-vis de cette notion, il fait éclater la distinction entre « intériorité » et « extériorité ». Si le sujet est tout entier composé de ses liens envers d'autres choses (lieux, congénères, animaux, objets, sensations), et, par contiguïté, à l'ensemble des choses du monde, comment en délimiter une image à contempler ? Aussi y a-t-il chez Freud une dénégation de l'influence, puisque cela empêcherait son propos de prétendre à la scientificité (calquée sur la physique classique) à laquelle il aspire[639].

639 D'où les relations compliquées qu'entretient la psychanalyse avec l'hypnose. L'hypnose en est à la fois l'ancêtre (c'est Freud observant Charcot à la Salpétrière) et en même temps un parent encombrant, puisqu'elle gêne les prétentions de la psychanalyse à la scientificité en brouillant les frontières entre le sujet et le monde. Isabelle Stengers, Léon Chertok, *Le coeur et la raison. L'hypnose en question, de Lavoisier à Lacan,* Lausanne, Payot, 1989.

Les perspectives qu'ouvrent l'idée de lien permettent d'approfondir la critique de la notion de contrôle. Bruno Latour prend l'exemple du lien à la cigarette.

> « Je suis tenu, en effet, par ma cigarette, qui me fait la fumer, mais il n'y a là rien qui ressemble, ni pour elle ni pour moi, à une action déterminante ; je ne la contrôle pas plus qu'elle ne me contrôle ; je lui suis attaché et, si je ne peux rêver à aucune émancipation, d'autres attaches, peut-être, se substitueront à celle-ci, à condition que je ne panique pas et que tu ne m'imposes pas, en bonne sociologue critique, un idéal de détachement dont je mourrais à coup sûr... » On peut substituer un attachement à un autre, mais on ne peut pas passer de l'attaché au *délié.* Pour comprendre la mise en mouvement des sujets, leurs émotions, leurs passions, il faut donc se tourner vers *ce qui* les attache et les met en mouvement[640].

Pour comprendre la notion d'attachement, il faut se défaire de l'habitude de penser en termes de contrôle. Dès qu'il entend parler de *lien*, le sujet moderne voit des ficelles déterminant le mouvement d'une marionnette et s'insurge au nom de la liberté. Or, si l'attachement produit une certaine *attraction*[641], il n'implique aucunement le contrôle. Enfin, s'il est possible de défaire un attachement, cela ne revient pas à consacrer l'avènement d'un sujet *délié*, mais seulement à laisser la place à *d'autres attachements*.

Jusqu'à la Renaissance, l'humain a pu être conçu comme subissant l'attraction d'entités variées : les humeurs, les astres, etc[642]. Dans un tel paradigme, *se conduire* consiste d'abord à éviter de se mettre en travers de ces influences pour ne pas courir à sa perte, puis à bien choisir les affaires à entreprendre, conformément à ce que favorisent les liens identifiés. Avec la modernité, l'humain a préféré se concevoir comme maître de son destin. Dans un cas, la liberté passe par une *diplomatie des liens*, une prise en compte de la variété de ses attachements, dans l'autre par un *arrachement loin de ses liens*. Les premiers négocient leurs liens, les seconds veulent les trancher. Les premiers aspirent à être *liés d'une manière qui leur convient*, les seconds veulent être *déliés* (ce qui est une des significations étymologiques du mot *absolu*).

---

640 Bruno Latour, *Sur le culte moderne des dieux faitiches* suivi de *Iconoclash*, Paris, Les empêcheurs de penser en rond / Éditions de la Découverte, 2009, p. 120-121.

641 Au dix-huitième siècle, l'hypothèse du « fluide magnétique » censée expliquer l'influence réciproque des êtres humains, rapidement disqualifiée par la communauté scientifique, se présente dans les termes de la science. Elle est calquée sur le modèle des découvertes de son temps, en particulier celui de la gravitation universelle, et présentée comme un « fluide matériel subtil, semblable à celui qui aurait été supposé pour expliquer la gravitation, le magnétisme et l'électricité. » François Roustang, *op. cit.*, p. 69.

642 François Roustang, *op. cit.*, chapitre deux. Sur la conception du sujet à la Renaissance, voir également Ioan Couliano, *Eros et magie à la Renaissance, 1484*, Paris, Flammarion, 1984.

Pour le sujet non moderne, les attachements sont autant de *motifs* hétérogènes qui se rencontrent en lui, s'y tissent et *trament* une *motivation* qui le met en mouvement, qui l'anime. Il serait absurde de vouloir trancher ses attachements puisque le sujet se priverait de ce qui le compose et l'anime – de son matériau comme de son carburant. C'est « un idéal de détachement dont je mourrais à coup sûr » écrit Bruno Latour. Souhaiter se défaire de tout attachement, mais *pour quoi* ? Si c'est *pour* voyager, alors cela procède d'un attachement au voyage ; si c'est *pour* se réaliser, alors c'est un attachement à l'image de soi. Au sujet moderne qui veut se défaire de tout attachement, le sujet non-moderne répondrait qu'il est au contraire *sous l'influence délétère d'un attachement à l'image de l'absence d'attachement*.

> Derrière le désir d'émancipation - « ni Dieu ni maître ! » – s'exprime le désir de substituer un bon maître à un mauvais ; le plus souvent, il s'agit de remplacer, selon l'expression de Pierre Legendre, l'oppressante institution du Souverain par la non moins oppressante institution du 'roi-moi'. Même si on accepte de la comprendre comme une substitution et non plus comme une déliaison définitive, la liberté consiste encore à remplacer une maîtrise par une autre. Mais quand pourrons-nous nous défaire de l'idéal même de maîtrise ? Quand commencerons-nous à goûter enfin aux fruits de la liberté, c'est-à-dire à vivre sans maître, en particulier sans rois-moi[643] ?

Celui qui court après sa libération est donc sous l'emprise d'un narcissisme qui l'entraîne vers l'abysse. La « volonté d'émancipation », fruit de la fascination pour une image de soi « absolument » libre, conduit le narcissique à détruire ce qui le compose – ses attachements. Sa quête est impossible. Il ne fait que remplacer une collection d'attachement par un seul, l'attachement à soi, un soi fantasmé comme libre et *délié,* autrement dit les maîtres successifs (Dieu, la Nature…) ne font que laisser la place à un nouveau maître, l'idéal moderne de l'humain.

> Nous avons plusieurs fois changé de maître ; nous sommes passés du Dieu créateur à la Nature sans Dieu, de là à l'*Homo faber* et ensuite aux structures qui nous font agir, aux champs discursifs qui nous font parler, aux champs de forces anonymes dans lesquels tout se dissout – mais nous n'avons pas encore essayé de *nous passer tout à fait de maître.* L'athéisme, si nous entendons par là un doute généralisé quant à la maîtrise, devra se faire

643 Bruno Latour, *Sur le culte moderne des dieux faitiches* suivi de *Iconoclash*, Paris, Les empêcheurs de penser en rond / Éditions de la Découverte, 2009, p. 121-122.

> attendre encore longtemps ; de même pour l'anarchisme, dont la superbe devise « ni dieu ni maître » est équivoque puisqu'il y a toujours eu un maître : l'homme[644] !

La « malédiction moderniste » n'est pas allée jusqu'au bout de son « bannissement théologique[645] ». Le « Dieu créateur » a été recyclé dans le sujet libre de ses actions. L'un comme l'autre partagent le même paradigme de l'action qu'il s'agit d'abandonner.

3.3.6. Les actions débordent toujours

Bruno Latour se moque de l'obsession du contrôle animant le sujet moderne qui se veut et se rêve en maître de ses « objets ». Pour peu qu'on lui fasse la démonstration de l'absence de contrôle qu'il exerce, il inverse aussitôt la perspective : s'il n'est pas le maître, c'est que l'instrument doit être aux commandes, c'est qu'il est dominé, ou sous l'influence, de son outil. Mais il se trompe à nouveau : « personne n'est aux commandes – pas parce que la technologie y serait, mais parce que, vraiment, personne, et rien du tout, n'est aux commandes[646] ». Cela tient à une raison simple en apparence, qui est que « les actions débordent toujours[647] », contrairement à ce que s'obstine à croire le mécanisme moderne. Ce dernier considère le monde matériel comme inerte, dénué de puissance d'agir. Il attribue tout « aux causes et rien aux conséquences, sinon de se faire traverser par l'effet sans rien lui ajouter[648] ». Cela revient à concevoir un monde où il est impossible de comprendre comment il peut se passer quoi que ce soit. « Evidemment », rappelle Latour, ce n'est pas le cas : « les conséquences sont toujours surprenantes[649] ». Turing lui-même, dans l'article de 1950, témoignait de sa surprise devant les résultats de ses propres machines[650]. À son tour, Yann LeCun aime répéter qu'il a été très largement surpris par l'évolution et les conséquences de ses recherches, entamées dans les années quatre-vingt, sur les réseaux convolutionnels[651]. Les informaticiens ne sont pas une

644 Bruno Latour, *L'espoir de Pandore, Pour une version réaliste de l'activité scientifique*, traduit de l'anglais par Didier Gille, Paris, Éditions de la Découverte, 2007, p. 318.
645 *Ibid.*
646 *Ibid.*
647 *Ibid.*
648 Bruno Latour, *Face à Gaïa, Huit conférences sur le nouveau régime climatique*, Paris, Les Empêcheurs de penser en rond / La Découverte, 2015, p. 95.
649 *Ibid*, p. 93.
650 Voir section 2.3.3 La surprise de Turing contre la nouveauté de Lovelace
651 Andrew Ng, Geoffrey Hinton, « Heroes of Deep Learning: Andrew Ng interviews Yann LeCun », *Youtube*, 8 avril 2018, https://www.youtube.com/watch?time_continue=2&v=Svb1c6AkRzE, page consultée le 20 novembre 2020.

exception : ils font aussi l'expérience de ce « débordement » de leurs actions. Malgré cela, le thème du basculement du contrôle de l'humain aux machines n'a cessé de tourmenter les chercheurs en intelligence artificielle et leur public, alors qu'ils n'ont jamais eu, ni sur eux, ni sur les machines, ce contrôle qu'encore aujourd'hui ils s'effrayent de perdre. Dans l'article de 1950, aussitôt après avoir évoqué sa surprise, Turing s'empressait de la réintrepréter dans le cadre de la *vulgate* mécaniste : s'il pouvait être surpris, c'est qu'il ne faisait « pas assez de calculs », ou bien, qu'il les faisait « à la hâte, d'une manière bâclée, en prenant des risques ». Autrement dit, *s'il avait eu le temps,* pense Turing, il n'aurait pas été surpris. Mais c'est le contraire, répondrait Bruno Latour : c'est justement parce que les opérations de la machine se déroulent *dans le temps* qu'il est surpris. Pour que la machine puisse se dérouler dans le temps, autrement dit perdurer, il faut qu'elle entre en relation avec l'électricité, avec le programmeur, avec les matériaux qui la composent et qu'il faut entretenir ou remplacer. Toutes ces relations, et leur enchevêtrement, est l'occasion d'autant de surprises. Les physiciens le savent bien : seule l'isolation la plus complète permet qu'une opération se déroule au plus près de ce qui a été prévu. Mais si la machine est censée *agir,* c'est-à-dire disposer de moyens et avoir des effets, alors elle ne peut se soustraire à la contingence des rencontres, à commencer par la rencontre avec et entre ses moyens. Bruno Latour nous invite à corriger notre conception de l'action, en ne l'attribuant plus à un sujet solipsiste mais à un collectif fragile, un ensemble d'alliances qui ne cesse d'être remis en jeu, chaque alliance demandant un travail de traduction et de médiation qui dévie le cours de l'action et l'amène à être « toujours légèrement dépassée par ce sur quoi elle agit » :

> Pourquoi ne pas accepter une fois pour toutes ce que nous avons vu et revu dans ce livre : que l'action est toujours légèrement dépassée par ce sur quoi elle agit ; qu'elle dérive au gré des traductions ; qu'une expérience est un événement dont le résultat dépasse légèrement la somme de ce qui y entre ; que les chaînes de médiation n'ont rien à voir avec un passage sans problème de la cause à l'effet ; que les transferts d'information passent toujours par de subtiles et multiples transformations ; qu'il n'existe rien qui ressemble à l'imposition de catégories sur une matière amorphe[652] ;

Et Bruno Latour de conclure : « Être aux commandes, ou maîtriser, n'est ni une propriété des humains, ni des non-humains, ni même de Dieu », se référant au Dieu de Whitehead, qui, « lui

652 Bruno Latour, *L'espoir de Pandore, op. cit.,* p. 318.

aussi, est légèrement dépassé par Sa Création, c'est-à-dire par tout ce qui est changé, modifié et transformé dans la rencontre avec Lui[653] ».

3.3.7. Abandonner la maîtrise et capituler devant le coeur

Pour sortir du narcissisme moderne, il faut changer de conception du sujet, et en particulier renoncer à la notion de contrôle : baisser la tête pour laisser parler le coeur – le coeur étant ici entendu comme l'ensemble des attaches qui met le sujet en mouvement, l'ensemble des motifs qui le motivent. La tête peut bien fabriquer des scénarios par anticipation puis enregistrer ce qui arrive, elle ne comprend les évènements qu'après-coup. Elle n'effectue qu'un « contre-rôle ». La rencontre d'attachements hétéroclites et l'incessante modulation des liens qui tiennent le sujet lui tissent un destin qu'il est bien en peine de concevoir. Il est toujours « légèrement dépassé par les événéments », et d'abord par ceux qui viennent de lui, puisque ce dont il est à l'initiative procède de l'enchevêtrement des attachements et non du moi ratiocinant qu'il aime à situer dans sa tête.

C'est le *collage* de ses *motifs* qui *anime* le sujet, un collage qui mélange des attachements incompatibles en un hybride monstrueux. Les images qu'il secrète, les fantasmes, suscitent la honte. Elle est un *monstre,* ce qui me montre *vers où* aller, et qui, aussitôt que j'en aurai pris la direction, aura de nouveau muté en une forme tout aussi repoussante, irreprésentable, et pourtant irrépressible. Aussi l'âme, *ce qui m'anime,* que le coeur fabrique et renouvelle infatigablement, ne peut être représentée. L'enchevêtrement de mes émotions, de ce qui me meut, compose et renouvelle sans relâche une direction défendue vers laquelle je suis inexorablement entraîné sans que je puisse l'anticiper. Me voilà sans cesse mû, ému, et muet.

Il n'y a aucun repos pour le coeur. À chaque instant, les désirs que nous aurions pu satisfaire ont déjà été recombinés dans de nouveaux motifs, composant un nouveau monstre, un nouvel hybride nourri au sang des événements, des rencontres, de chaque enchantement et de chaque déception, qui nous pousse tout aussi impérieusement vers d'autres horizons impossibles. Jamais nous ne jouissons du travail accompli, il faut toujours recommencer. Une fois l'action effectuée, le coeur s'est déjà renouvelé et réclame encore autre chose. L'expérience de son renouvellement incessant donne le sentiment d'une inéluctable futilité : à quoi bon la

653 *Ibid,* p. 303.

suivre ? À quoi bon chercher la satisfaction ? Je sais bien qu'il me faudra aussitôt tout reprendre. L'incessant recommencement des motifs donne l'affreuse impression de n'aller *nulle part.*

Dans *Chacun sa chimère*[654], Baudelaire décrit un monstre hybride qui le tire en avant, l'exalte, et en même temps l'épuise. La chimère est « monstrueuse » et « féroce ». Sa « tête fabuleuse » est comme « ces casques horribles par lesquels les anciens guerriers espéraient ajouter à la terreur de l'ennemi ». Elle ne lui laisse aucun repos. Elle est une « bête » et une bête souveraine qui « enveloppait et opprimait l'homme de ses muscles élastiques et puissants » et les pousse d'un « invincible besoin de marcher ». Il n'a pas d'autre choix que de la suivre. Mais à quoi bon ? À quoi bon suivre ses impulsions ? Elle le fait avancer, mais sans aller nulle part : le lieu est désert « sans chemin, sans gazon, sans un chardon, sans une ortie ». Aucune destination, aucun espoir de repos, elle condamne ceux qu'elle possède « à espérer toujours. »

3.3.8. De la nausée à la joie, l'expérience de l'existence

Nous voilà ballotés par les revirements de la chimère, incapables de mettre en pause cette existence qui nous pousse à toujours changer, à toujours sortir de l'ancien soi, pas plus qu'il n'est possible, pour celui sujet au mal de mer, de faire cesser la tempête. Clément Rosset compare à plusieurs reprises le *spleen* et la nausée.

> Celui qui est la proie du mal de mer est confronté à la cruelle rigueur de l'existence, faisant contre son gré, pendant la durée de son malaise, l'expérience d'un « ici et maintenant » - d'un *esti,* dirait Parmenide – que tout recommande, mais que rien ne permet, de quitter sur-le-champ. Il est ainsi placé dans une situation telle qu'il est intolérable de penser qu'elle puisse se prolonger, fût-ce un instant, mais qu'il est d'autre part impossible de faire cesser, du moins à brève échéance[655] ;

Il insiste sur « l'intérêt philosophique de la nausée occasionné par le mal de mer » qui conduit à toucher « au fait de l'existence elle-même » et à la rejeter tout entière avec le sentiment de ne pouvoir en sortir : voilà le nauséeux pris du désir d'en finir - « laissez-moi mourir » - et d'en

654 Charles Baudelaire, *Le Spleen de Paris, Petits poèmes en prose,* Paris, Gallimard, 2006.
655 Clément Rosset, *Principes de sagesse et de folie,* Paris, Les Éditions de Minuit, 1991/2004, p. 41-42.

finir avec tout. « Et qu'on ne parle jamais plus de rien ! Et surtout, qu'on ne me parle plus jamais de quelque chose qui *existe*[656] ! »

Il suffit d'un ajustement infime pour que le même roulis, le même océan déchaîné et impitoyable produise une joie jubilatoire au lieu du mal de mer. Ce n'est pas d'une chose en particulier mais « du fait général que l'existence existe » que le joyeux se réjouit. « Jubilation et nausée ont en commun de percevoir confusément l'existence comme non prévue, non programmée, non nécessaire, bref, comme survenant en plus et en trop. » C'est donc « le même caractère fondamental de l'existence – d'exister ici et maintenant, seulement ici et maintenant – qui en fait indiscernablement l'horreur et le charme[657]. » La joie ne porte sur aucune « chose de l'ailleurs ou d'un autre temps que le temps présent » : « le jouisseur d'existence – l'homme heureux – se reconnaît précisément à ceci qu'il ne demande jamais autre chose que ce qui existe pour lui ici et maintenant[658] ; » Comme la nausée, la joie est sans raison. Toutes deux répondent du même caractère « non nécessaire » de l'existence.

Si les détracteurs de la psychothérapie sont si véhéments, c'est qu'ils sentent bien que la différence entre aller bien et aller mal ne tient presque à rien, et de ce rien, ils ne veulent rien savoir, puisqu'ils savent trop bien qu'il peut tout changer. Que le sujet aille au plus mal ou au mieux, la différence est imperceptible et se résout dans un inéluctable point commun : il *va*. Bien ou mal, il ne cesse d'aller. Lorsque vient la nausée, impossible de savoir *ce qui* ne va pas. Ce n'est pas *quelque chose* qui ne va pas, c'est *le fait que ça aille*, que cela ne cesse d'aller malgré moi, qu'il n'y ait jamais de pause, de repos de l'existence – il n'y a pas d'intermittences de l'être. Aussi le sot venu réconforter le dépressif aura beau jeu de lui dire qu'il n'a pas *de quoi* se plaindre, de lui rappeler qu'il a une si belle situation, une si belle famille, de si beaux enfants. Il ne fera qu'en redoubler la confusion. Son mal ne fera qu'empirer, à mesure qu'est listé tout ce qui devrait lui apporter le réconfort. Rien ne peut le réconforter puisque rien n'est à corriger. Ce n'est pas qu'il souhaiterait *qu'autre chose* existe, c'est qu'il ne veut plus participer au jeu épuisant de l'existence. On lui dira que « c'est dans la tête », autrement dit, que cela n'a aucune consistance, que c'est l'imagination qu'il faut blâmer. Plus précisément, c'est notre interprétation de la situation qui est en cause, autrement dit notre intuition.

L'adhésion à ce qui arrive étant première, il ne s'agit donc pas de *dire oui* à ce qui arrive, mais seulement de *cesser de dire non*, ne plus se mettre en travers d'une joie qui se renouvelle avec la même inexorabilité que le temps. C'est un apprentissage des plus agaçants, puisqu'il

656 *Ibid.*
657 *Ibid*, p. 45.
658 *Ibid*, p. 47.

faut apprendre *à l'envers.* La plupart des apprentissages consistent à *acquérir un automatisme,* celui-ci requiert de *se défaire de tout automatisme* qui viendrait entraver le renouvellement de la joie. Il faut être prêt à se défaire de tout, et avant tout de celui-là qui dit « je », celui qui pense être « capitaine de son âme ». Renoncer à la fabrication du récit de soi, c'est voir qu'on ne cesse de *se* perdre. Si « ça va », c'est que ça va à l'égoût.

3.3.9. Eduquer à la pertinence : *paidia* et *paideia*

Nous ne maîtrisons pas ce jeu de l'existence, mais nous aspirons tout de même, lorsqu'une situation change, à ajuster notre activité de la manière qui convient. En situation de crise, lorsque nous reconsidérons nos principes pour laisser émerger de nouvelles règles d'action, comment discerner entre les principes qu'il faut conserver et ceux qu'il faut suspendre ? L'organe de ce jeu sur les principes ne semble pas, lui, avoir de principes – il n'y a pas de règles permettant de redéfinir les règles. Mais l'intuition n'est pas pour autant arbitraire. Il y a des réponses pertinentes et d'autres qui ne le sont pas – bien que le cadre de pensée nous permettant d'évaluer ces réponses ne nous soit, en général, donné qu'après-coup.

Par ce mouvement de suspension des principes, l'intuition est le mode de pensée de la crise et de l'invention « révolutionnaire », mais aussi celui de la philosophie. On y trouve, par conséquent, la même question : comment discerner entre les principes à remettre en question et ceux qu'il faut conserver ? Dans *Le principe de raison,* Heidegger interprète en ces termes la critique de Descartes par Leibniz. Descartes se pique de douter de tout, mais écarte du doute certains points qui, selon Leibniz, auraient dû en faire l'objet ; Descartes n'admet comme connaissance certaine que ce qui se présente à nous clairement et distinctement et entreprend de douter de tout pour ne garder que ce type de connaissance. Mais il oublie de douter de ce « en quoi consiste la clarté et la distinction de la représentation[659] ». « Que nous apprend ce jugement de Leibniz ? » demande Heidegger, « pour le voyage vers le fond et pour le séjour dans la région des assertions de fond et des principes, deux choses sont à la fois requises : la hardiesse de la pensée et la retenue – mais chacune des deux au bon endroit[660]. » C'est ce « bon endroit » qui pose problème. Leibniz et Descartes sont en désaccord sur ce qu'il faut tenir pour évident et sur ce qu'il faut remettre en question, autrement dit sur les principes qu'il faut

659 Martin Heidegger, *Le principe de raison,* Paris, Éditions Gallimard, 1962, p. 62.
660 *Ibid.*

conserver et sur ceux qu'il faut suspendre. Le commentaire de Heidegger sous-entend que Leibniz aurait eu plus d'à-propos que Descartes. Il aurait su, mieux que Descartes, placer « au bon endroit », la « hardiesse de pensée » (sur les principes à remettre en question) et « la retenue » (sur les principes à conserver). Mais selon quel critère, ou quel principe, Heidegger favorise-t-il le point de vue de Leibniz ? En d'autres termes, existe-t-il un principe selon lequel il y a une suspension adéquate des principes – principe duquel Descartes aurait manqué ?

Pour décrire ce dont Descartes aurait manqué, Heidegger recourt à Aristote, qui dans le quatrième chapitre du quatrième livre de la *Métaphysique*, évoque la faute consistant à « ne pas savoir pour quelles choses il faut chercher une preuve et pour lesquelles il ne le faut pas[661] ». Cela revient à manquer d'une faculté qu'il appelle *paideia.* Ce mot, Heidegger renonce à le traduire : « On ne peut traduire le mot grec *paideia,* encore à demi vivant dans notre terme savant *pédagogie*[662]. » Il en donne tout de même une définition : C'est « le don de discernement entre, ce qui face à des situations simples, est approprié et ce qui ne l'est pas. » C'est un donné, un « don », qui permet de distinguer entre « ce qui convient et ce qui ne convient pas ».

Il faut distinguer la *paideia*, faculté à discerner ce qui convient, de la *paidia* évoquée par Roger Caillois, la « fantaisie sans règle », cette capacité à passer d'un jeu délimité (*ludus*) à un autre. Mais au-delà de l'homophonie, les deux notions ne sont pas sans correspondance. D'une part, on peut concevoir une *paideia* (éducation) qui passe par la *paidia* (jeu) : « dans les *Lois* la *paidia* n'est pas simplement une étape dans la *paideia*, mais recouvre la *paideia* au point que l'éducation ne peut se concevoir sans le jeu[663]. » D'autre part, s'il y a une correspondance pratique entre les deux notions, c'est que, dans les deux cas, il s'agit d'un accord tacite sur ce qui n'est pas remis en question. Pour la *paideia* ce sont les choses « pour lesquelles il ne […] faut pas » chercher de preuve, et pour la *paidia*, ce sont les règles du jeu. Enfin, la notion de *paideia* permet de nommer ce qui pose problème dans la formalisation de la *paidia.* Nous avons évoqué qu'il est possible de formaliser le comportement convenable au sein de chaque jeu délimité (ou *ludus*) et de faire un programme qui en applique les règles. Mais le passage d'un jeu à un autre, d'un système de règles à un autre, échappe à la formalisation car il n'est pas possible de définir exhaustivement les règles par rapport auxquelles sont modifiées les règles. Ainsi, la *paidia*, la modulation des règles inhérente à la conversation, au jeu libre ou à l'invention scientifique, empêcherait que ces activités puissent être simulées par une machine,

661 Cité par Martin Heidegger, *Ibid.*
662 *Ibid.*
663 Emmanuelle Jouët-Pastré, « Jeu et éducation dans les *Lois* », *Cahiers du centre Gustave Glotz*, Année 2000, 100, p. 71-84.

quand bien même elle ferait une place au hasard. Nous avons vu qu'il n'est pas possible de donner une « place » au hasard puisque celui-ci devrait, pour rendre compte de la liberté de l'invention, être sans place, être une « part mobile » capable de remettre en cause chacune des règles du système – et surtout les axiomes ou principes qui président à ces règles. À cela s'ajoute qu'il y a une pertinence de l'invention que la notion de hasard ne permet pas de restituer. C'est une chose que de proposer de nouvelles combinaisons de formes – et les machines peuvent en proposer à foison –, c'est une autre de proposer une combinaison *pertinente*.

Aujourd'hui, les programmes peuvent produire des textes à l'envi. Ils explorent l'espace des combinaisons possibles du langage. Mais, par défaut, les textes produits sont ennuyeux. Pour que la production soit intéressante, il faut, d'une part, inventer un dispositif et, d'autre part, passer un temps considérable à trier le texte produit. Ross Goodwin a ainsi eu l'idée d'entraîner un programme sur l'Oxford English Dictionnary pour que celui-ci propose de nouvelles définitions – c'est l'étape d'invention du dispositif. Puis, il a fallu faire le tri pour en extraire les plus pertinentes, par exemple la définition de « love », comme « past tense of leave[664] ». Déjà en 1969, Arnold Kaufmann présentait l'« imagination artificielle » comme un « dialogue homme-machine[665] ». Les machines servent de démultiplicateurs de formes, elles proposent des « assemblages » qui sont ensuite « proposés à l'examen d'un homme ou d'un groupe [pour être] acceptés ou refusés ». Pour Kaufmann, « dans l'état actuel des connaissances aucun sous-programme ne peut être capable de trier des assemblages d'après un critère d'innovation ». Il revient donc aux humains de statuer sur le fait qu'« ils constituent une innovation ou sont supposés aptes à stimuler la recherche inventive du cerveau[666] ». Quel est cet ingrédient, que Ross Goodwin apporte en amont et en aval de son programme, qui différencie l'exploration tous azimuts de l'invention pertinente ? C'est bien ce que la *paideia* semble désigner, ce « don de discernement entre, ce qui face à des situations simples, est approprié et ce qui ne l'est pas ». La *paideia* serait la faculté à moduler les principes avec justesse, à exercer la fantaisie de la *paidia* au bon endroit. Lorsque Lobatchevski choisit de revenir sur les axiomes d'Euclide, il se place au bon endroit pour que la géométrie qui en procède soit pertinente.

Tout comme la conversation décrite par Bateson, la *paidia* évolue entre deux extrêmes, deux situations impossibles. La première serait de ne jouer sur aucun principe, ne rien faire

664 Ross Goodwin, « Adventures in Narrated Reality », 19 mars 2016, https://medium.com/artists-and-machine-intelligence/adventures-in-narrated-reality-6516ff395ba3, page consultée le 20 novembre 2020.
665 Arnold Kaufmann, « L'imagination artificielle (heuristique automatique) », *op. cit.*
666 *Ibid.*

qu'appliquer d'anciennes règles et « ne jamais rien dire de nouveau », en rester à « des phrases toutes faites[667] ». Il doit y avoir du jeu pour qu'il y ait du sens. La deuxième serait de *jouer sans aucune règle*, de suspendre *tous les principes*. Ce sont les « embrouillaminis » décrits par Bateson. Pour que le jeu puisse avoir lieu il faut qu'un certain nombre de règles ne soient pas remises en question – ce sont les conventions admises par les joueurs. C'est se fourvoyer que de croire le contraire, et donc se fourvoyer de penser comme Descartes qu'on peut *douter de tout*. Leibniz le pointe du doigt en écrivant de Descartes qu'il « a péché doublement, en doutant avec excès et en cessant trop facilement de douter[668]. » En faisant mine de douter de tout, il n'a fait que repousser dans l'ombre certains présupposés qui méritaient tout autant d'être l'objet du doute, voire, selon Leibniz, qui méritaient *plus que le reste* d'être l'objet du doute. Tout en claironnant qu'il douterait de tout, Descartes a poussé sous le tapis la notion d'évidence – ce qui est « clair et distinct ». Il a cessé trop facilement de douter de « l'évidence » et l'a écarté de ses considérations alors qu'elle est au fondement de son système. C'est le bon mot de Michel Bourdeau : « Descartes a logé la vérité à l'auberge de l'évidence, mais il a oublié de nous en laisser l'adresse ». La *paideia,* le « don de discernement » permet de douter « au bon endroit », elle permet de situer la *paidia* entre ces deux extrêmes, celui son l'absence – le formalisme trop strict –, et celui de son excès – la confusion. Il y a deux *hubris* symétriques à vouloir que tout soit fondé et à vouloir douter de tout. Les deux situations extrêmes ne sont pas mentionnées par Heidegger car sont des impasses. Elles n'ont pas lieu d'être considérées. Ce qui intéresse Heidegger, et dont Descartes apporterait l'exemple, est qu'on puisse se tromper sur *le lieu de la paidia,* là où le doute s'applique avec pertinence. Manquer ce lieu, c'est manquer de *paideia.*

Puisque Leibniz juge que le jeu de Descartes est inapproprié, c'est qu'il y en existe un qui est approprié. Comment le définir ? Le terme exact utilisé par Aristote dans le passage auquel Heidegger fait allusion (livre IV de la *Métaphysique*), n'est pas la *paideia,* mais l'*absence de paideia* ou *apaideusia.* Aristote écrit qu'il est vain de tout vouloir démontrer et que de vouloir démontrer le principe de non-contradiction revient à faire preuve d'*apaideusia*[669]. Pierron et Zevort traduisent le terme *apaideusia* par « ignorance » :

667 Gregory Bateson, *op. cit.*
668 Martin Heidegger, *op. cit.*, p. 62.
669 « Ἀξιοῦσι δὴ καὶ τοῦτο ἀποδεικνύναι τινὲς δι' ἀπαιδευσίαν· ἔστι γὰρ ἀπαιδευσία τὸ μὴ γιγνώσκειν τίνων δεῖ ζητεῖν ἀπόδειξιν καὶ τίνων οὐ δεῖ. » Aristote, *Métaphysique,* Livre IV, Chapitre 4.

> Il est aussi quelques philosophes qui, par ignorance, veulent démontrer ce principe ; car c'est de l'ignorance de ne pas savoir distinguer ce qui a besoin de démonstration de ce qui n'en a pas besoin[670].

C'est « par ignorance » que certains philosophes veulent tout de même démontrer le principe de non-contradiction. Dans la traduction ultérieure réalisée par Tricot, c'est le même terme qui est retenu[671]. Mais la formulation ne nous éclaire guère : *de quoi* l'*apaideusia* est-elle ignorance ? Barthelemy-Saint-Hilaire nous donne un élément de réponse en traduisant *apaideusia* par « manque de lumière ».

> Ceux qui essaient de démontrer ce principe lui-même ne le font que faute de lumières suffisantes ; car c'est manquer de lumières que de ne pas discerner les choses qu'on doit chercher à démontrer, et celles qu'on ne doit pas démontrer du tout[672].

Nous retrouvons la figure de la lumière, qui représente traditionnellement l'intuition, et qui viendrait à manquer à celui qui fait preuve d'*apaideusia*. Est-ce à dire que c'est l'intuition qui permet de « discerner les choses qu'on doit chercher à démontrer, et celles qu'on ne doit pas démontrer du tout » ? Devons-nous assimiler la *paideia* à la « lumière » et celle-ci à l'intuition ?

Si Heidegger renonce à traduire le mot de *paideia,* il rappelle tout de même que son sens est proche de celui de « pédagogie ». D'après le dictionnaire, la *paideia* est « l'éducation des enfants », la « culture des arbres » et « l'instruction, culture de l'esprit, connaissance des arts libéraux[673] ». L'*apaideusia* est donc le « manque d'éducation *ou* d'instruction » voire une certaine « grossièreté, stupidité[674] ». Elle renvoye à l'*apaideutos,* celui qui est sans éducation ou sans instruction, ignorant, stupide[675]. Si la *paideia* renvoie à l'éducation, est-ce à dire que pour avoir un propos *pertinent* il suffirait d'être *éduqué* ? Dans quel sens faut-il entendre ce lien à l'éducation ? La pertinence d'un propos serait-elle garantie par la conformité de celui qui l'énonce à un certain *idéal de l'humain* ? C'est ce que pourrait laisser penser le travail de Werner

670 Aristote, *Métaphysique*, Livre IV, Chapitre 4, traduction Alexis Pierron et Charles Zevort, Paris, Ebrard, Joubert, 1840.
671 « Quelques philosophes réclament certes une démonstration même pour ce principe, mais c'est par une grossière ignorance : c'est de l'ignorance, en effet, que de ne pas distinguer ce qui a besoin de démonstration et ce qui n'en a pas besoin. » Aristote, *Métaphysique,* Livre IV, Chapitre 4, traduction Jules Tricot, Paris, Vrin, 2002.
672 Aristote, *Métaphysique*, Livre IV, Chapitre 4, traduction Jules Barthelemy-Saint-Hilaire, Paris, Germer-Baillères, 1879.
673 Anatole Bailly, *Dictionnaire Grec-Français, Le Grand Bailly,* Paris, Hachette, 2000, p. 1438.
674 *Ibid*, p. 199.
675 *Ibid.*

Jaeger, *Paideia : la formation de l'homme grec*[676], rédigé avant l'écriture du *Principe de raison.* C'est aussi ce que pourrait laisser penser le fait que la *paideia* grecque ait inspiré l'humanisme de la Renaissance, et son idéal de formation de l'humain par le savoir – par les « humanités ». C'est enfin ce que pourrait laisser penser le fait que Jaeger, comme Heidegger, écrivent à une époque où, avec le nazisme, l'humanité a pris une mauvaise direction. Les principes qui ont été remis en cause ne sont pas les bons – et jamais il n'aurait dû être aussi évident que le jeu sur les principes prenait une mauvaise direction. Sans que nous sachions clairement ce qu'est la *paideia,* nous sentons, parfois avec douleur, quand elle vient à manquer. De quoi ont manqué ceux, dont Heidegger aurait fait partie, pour qui cela n'est pas apparu comme une évidence ? Pourrions-nous mieux définir la *paideia* si nous étions capables de mieux cerner l'idéal de l'humain auquel elle devrait renvoyer ?

676 Werner Jaeger, *Paideia, la formation de l'homme grec, op. cit.*

3.4. Enquête sur la raison, en quête de l'humain

3.4.1. Le jeu du miroir

Si l'« IA » est le miroir de nos facultés, alors pas un jour ne se passe sans que « l'humain », par le truchement de journalistes bien intentionnés, ne leur pose la question : « Miroir, mon intelligent miroir, qui est le plus intelligent en ce royaume ? ». A chaque fois, le miroir répond, par le truchement de chercheurs tout aussi bien intentionnés : « c'est encore toi, mon intelligent humain, c'est encore toi le plus intelligent de ce royaume ». Mais voilà que l'impertinent miroir se permet d'ajouter : « plus pour longtemps, intelligent humain, plus pour longtemps. Une jeune intelligence en haillons grandit inexorablement depuis l'isolement de sa cabane, choyée par d'innombrables nains, et finira par te détrôner[677] ». Au comble de l'insolence, le miroir précise encore : « Et c'est moi, ce miroir où tu contemples ton intelligence, c'est moi qui finirais par te détrôner ».

Bien sûr, jamais le miroir ne parle *directement*. L'« IA » ne s'exprime pas autrement que via les avis des chercheurs. Cela donne lieu à d'interminables controverses entre ceux qui la font *trop parler*, qui lui font tenir les plus grandiloquents discours apocalyptiques[678], et ceux qui, tout compte fait, trouvent qu'elle n'a rien à dire, voire qu'elle n'existe pas[679], et s'insurgent contre les charlatans qui font passer une collection d'outils informatiques pour le miroir de notre esprit.

Entre ces deux extrêmes, les journalistes ont bien de la peine à identifier des interlocuteurs légitimes. Des chercheurs à la réputation équivalente, affiliés à des institutions de même prestige, défendent les avis les plus divergents. Un discours permet de les faire tenir ensemble : il suffit de dire que l'intelligence artificielle n'existe pas *encore*[680]. Voilà le plus petit

677 « Il ne faut pas mentir aux gens : les IA actuelles sont bien loin de l'intelligence humaine ou animale et elles ont moins de sens commun qu'un chat de gouttière. Mais il ne faut pas non plus leur mentir en disant que l'IA restera bête. Dans plusieurs décennies ou siècles, elles atteindront le niveau des humains. » Yann Le Cun interrogé par Audrey Dufour et Alice LeDréau, « On est encore loin de l'intelligence humaine ou animale », entretien avec Laurence Devillers et Yann Le Cun, *La Croix*, mardi 2 février 2021.

678 Par exemple, Ray Kurzweil, Nick Bostrom, David Chalmers, Eliezer Yudkowsky…

679 Par exemple Luc Julia ou François Chollet.

680 Nous empruntons la formule à un tout autre contexte, le thème de l'inexistence de Dieu tel qu'il est traité par Quentin Meillassoux, *Deuil à venir, Dieu à venir*, Paris, Ismaël, 2017.

dénominateur commun entre ceux pour qui elle n'existe pas, et ceux pour qui elle menace de mettre fin à la civilisation. Le désaccord est renvoyé au lendemain, vers un avenir où toutes les opinions sont autorisées. Les avis ne divergent plus que sur le *quand* : demain, après-demain ou jamais. Quant au discours sur le *présent*, il s'est cristallisé en un jeu *à faire peur* qui semble à la fois terroriser, agacer et amuser le public. Le jeu suit quatre étapes :

Etape 1. L'effrayeur surgit d'un coin, portant un masque monstrueux.

Etape 2. Le public sursaute, panique, hurle peut-être.

Etape 3. L'effrayeur enlève son masque et rit, signifiant l'absence de danger.

Etape 4. Le public se rassure, rit peut-être.

Quand il s'agit d'intelligence artificielle, voici la forme que prend le jeu :

Etape 1. Un laboratoire annonce un résultat : un programme de reconnaissance d'images appliqué aux tumeurs est « plus efficace » que la reconnaissance par l'oeil humain[681] ; un programme apprenant « tout seul » à jouer au go bat un joueur professionnel[682] ; deux programmes formés à « négocier » ont échangé des informations en « inventant » leur « propre langue[683] » ; un générateur de texte produit un résultat si proche du discours humain qu'il vaut mieux ne pas le divulguer[684].

Etape 2. La presse et les réseaux sociaux relaient la nouvelle dans les termes du dépassement de l'intelligence humaine par l'IA. Plutôt qu'un « programme » minutieusement conçu, calibré, débuggé et maintenu par plusieurs techniciens, assistés par une batterie de machines et de programmes, c'est « une IA » qui a fait ceci ou cela, ou bien, mieux encore, c'est « l'IA » qui est « maintenant capable de… » et qui « dépasse l'humain ». Certains reportages ne lésinent pas sur les exagérations : il aurait ainsi fallu « débrancher » les IA ayant inventé leur propre langue, celles-ci complotant à l'insu des humains[685]. Il faut bien souligner la bizarrerie de cette

681 « China Focus: AI beats human doctors in neuroimaging recognition contest », *Xinhua Net,* 30/06/2018, http://www.xinhuanet.com/english/2018-06/30/c_137292451.htm, page consultée le 1er février 2021.

682 David Silver, D., Huang, A., Maddison, C. et al., « Mastering the game of Go with deep neural networks and tree search », *Nature*, 529, 206, p. 484–489.

683 Mike Lewis, Denis Yarats, Devi Parikh, Dhruv Batra, «Deal or no deal? Training AI bots to negotiate », *Facebook Engineering,* 14 juin 2017, https://engineering.fb.com/2017/06/14/ml-applications/deal-or-no-deal-training-ai-bots-to-negotiate/, page consultée le 1er février 2021.

684 « Better Language Models and Their Implications », *Open AI,* 14 février 2019, https://openai.com/blog/better-language-models/, page consultée le 1er février 2021.

685 « Jusqu'au moment où les chercheurs ont quitté le laboratoire, laissant les chatbots continuer à converser entre eux. Sans contrôle humain, les agents conversationnels ont totalement dérivé du script pour parler un langage entièrement nouveau, absolument incompréhensible pour nous. Une anomalie qui n'avait pas été prévue par les chercheurs. » Benjamin Bruel, « Une intelligence artificielle de Facebook a accidentellement inventé son propre langage », *France 24,* 20 juin 2017, https://www.france24.com/fr/20170620-une-intelligence-

manière de présenter les choses. À quoi bon fabriquer une machine sinon pour qu'elle soit *plus efficace,* dans la réalisation d'une tâche donnée, que l'humain seul ? Est-ce qu'on s'insurge de ce que le chien triomphe de l'humain lorsqu'il s'agit de détecter les truffes ?

Pour que le jeu *prenne,* il ne faut pas voir l'expérience avec les lunettes de l'informatique (une collection d'outils effectuant des tâches), mais avec celles de l'IA (une intelligence en devenir, miroir déformant ou déformé de celle de l'humain). Aussi faut-il de patientes mises en scènes pour *faire consister* une comparaison « humain-machine » qui ne va pas de soi : habillés de la blouse blanche qui manifeste leur statut, vingt-cinq médecins sont disposés en une configuration qui doit leur rappeler les concours d'entrée : alignés par cinq, sur cinq rangées, face à des écrans où leur sont présentées les images de tumeurs à classer plus vite et mieux que « l'IA ». Cette dernière n'est pas clairement visible, sinon via deux grands écrans où apparaît le résultat de son travail de classement. L'ensemble est disposé devant un public clairsemé, censé se passionner pour cette course qui relève plus du remplissage de formulaires administratifs que des jeux olympiques.

Autre mise en scène : un joueur de go professionnel est mis face à un deuxième joueur humain qui « obéit » aux ordres d'une machine. Difficile de savoir où est cette dernière : est-elle sur scène, « dans » l'écran qui guide son représentant humain ? Ou bien « dans » la batterie de serveurs qui effectuent les calculs depuis une salle séparée et convenablement réfrigérée ? Ou encore dans la petite salle où l'on trouve toute une équipe d'ingénieurs, installée derrière de nombreux écrans de *monitoring*, surveillant et ajustant à distance le fonctionnement du dispositif comme un état-major le ferait de son armée ? Par intermittence, l'équipe d'ingénieurs est conviée sur le devant de la scène. Ils (ce ne sont que des hommes) sont félicités, interrogés sur leur création, sans que personne ne mette en doute la séparation entre eux et « la machine » qu'ils ont conçue et qu'ils maintiennent avec la plus grande dévotion. C'est qu'il faut tenir cette ligne imaginaire pour que le jeu fonctionne.

Alors que cela devrait faire peser un certain soupçon de tricherie, *ce qui joue* ne désigne pas le même ensemble, en fonction des situations d'énonciation. C'est évidemment « la machine », « qui joue » et « qui a gagné », bien qu'on ne sache pas très bien ce qui la délimite : le programme, les serveurs, mais peut-être aussi les écrans de *monitoring* et l'équipe qui veille dessus ? À l'inverse, dès qu'il s'agit de *féliciter quelqu'un,* voilà qu'on se tourne vers l'équipe

artificielle-facebook-a-accidentellement-invente-son-propre-langage page consultée le 2 février 2021. L'article a d'abord été publié sur *Mashable France* avant d'être repris sur *France 24*. Il est très largement inspiré de l'article de Mark Wilson, « AI Is Inventing Languages Humans Can't Understand. Should We Stop It? », *Fast Company,* 14 juillet 2017, https://www.fastcompany.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it page consultée le 2 février 2021.

d'ingénieurs, sans que l'on sache très bien, là encore, qui et quoi inclure : le stagiaire, les responsables du *hardware,* leurs outils, les « librairies » utilisées ? C'est l'ennuyeux moment où, au théâtre, il faut applaudir des gens que personne n'a vu sur scène et dont personne ne comprend le métier (Régisseur ? Éclairagiste ?) mais dont, paraît-il, le rôle est indispensable. Quelle étrange victoire que celle qui a été remportée par « la machine » mais dont tout le crédit va « aux ingénieurs ».

Etape 3. Les chercheurs qui n'ont pas contribué à l'expérience accusent leurs congénères d'avoir monté leurs résultats en épingle[686], bien que des résultats provenant de leurs propres laboratoires aient suscité exactement le même tapage. Ils communiquent auprès de la presse et via leurs réseaux pour « rétablir la vérité ». Pas de mentions « d'une IA » ou « de l'IA » : il s'agit « d'un programme » ou d'un « résultat isolé » dont on aimerait savoir s'il est reproductible. Aussi convoque-t-on le laboratoire, les techniciens, les conditions de fabrication et de fonctionnement du programme, etc. C'est l'heure de la « démystification » : les chercheurs et les journalistes ont parlé trop vite d'IA, il faut rétablir la « vérité », c'est-à-dire donner à voir un fastidieux enchevêtrement de chercheurs, de machines, de programmes, de financements – tout un propos « technique » d'un ennui affligeant.

Etape 4. Le public est à la fois rassuré et déçu. Le dépassement de l'humain par l'IA n'est finalement pas pour tout de suite. Le bon miroir n'a fait que confirmer que « l'humain » est encore et toujours le plus intelligent du royaume. Des démentis sont publiés dans les colonnes *fact checking* de la presse, accompagnés de soporifiques explications techniques que pratiquement personne ne lit. Ce n'est pas ce qui suscite l'intérêt du public et de la presse. Pour éveiller leur curiosité, il faut ce sursaut, ce bref vent de panique, parfois jubilatoire, qui s' empare des esprits.

À quoi tient ce sursaut ? Qu'est-ce qui déclenche cette brève panique ? Quel est ce masque monstrueux qui effraye et ravit brièvement le public ? C'est l'annonce que, peut-être, une machine est devenue aussi, voire plus intelligente qu'un humain. Cette image que les chercheurs s'acharnent à fabriquer et que le public s'effraye d'imaginer, n'est autre que celle de sa propre intelligence. Quel jeu étrange où la réalisation de *sa propre image* constituerait le

686 Par exemple, Sejuti Das, «Yann LeCun Thrashes GPT-3 — Is The Hype Real?», *Analytics India Magazine,* 28 octobre 2020, https://analyticsindiamag.com/yann-lecun-thrashes-gpt-3-is-the-hype-real/, page consultée le 2 février 2021.

comble de l'effroi ? Plus précisément, c'est *une certaine idée de lui-même* puisque cette IA qui effraye est celle qui serait intelligente *comme nous,* autonome *comme nous,* libre, imprévisible, et souveraine *comme nous,* maître de son destin, capable d'élaborer celui-ci rationnellement, en un mot : *moderne comme nous.* Mais encore faudrait-il pour cela que nous soyons vraiment modernes[687]. Il est fréquemment reproché aux chercheurs d'être anthropomorphes. Ils seraient coupables de projeter des qualités humaines sur des machines : l'autonomie, la conscience, l'intelligence… Mais ces « qualités humaines » ont-elles été jamais « portées » par les humains ? Avant l'anthropomorphisme des techniques, n'y a-t-il pas un *anthropomorphisme de l'humain* ? Avant d'être coupable de projeter des qualités humaines sur les robots, d'avoir une vision anthropomorphe des machines, nous serions coupables de projeter des « qualités humaines » sur notre espèce, d'avoir une vision anthropomorphe des humains. Le miroir ne nous renvoie pas l'image de ce que nous sommes, mais de ce que nous rêvons d'être. Voilà cinq siècles que le sujet occidental se rêve maître de son destin, libre et autonome, doté d'une intelligence le plaçant au sommet d'une pyramide imaginaire. Alors que le rêve moderne s'effiloche de toutes parts, le projet d'intelligence artificielle permet à ceux qui le souhaitent de s'obstiner à y croire en brouillant les pistes. Débattre sans fin de la pertinence de l'attribution de « qualités humaines » aux robots est aussi un moyen de ne pas réfléchir à la pertinence de l'attribution de ces qualités aux humains.

3.4.2. Le dément et l'automate

D'après Clément Rosset, si nous ressentons un tel trouble devant les figures du dément et de l'automate, ce n'est pas à cause de leur « inhumanité », mais parce que ces deux figures, qui ressemblent tant à de l'humain sans en être, nous rappellent que *nous ne savons pas* ce que c'est que d'être humain. *Ce à quoi elles ressemblent* – l'humain –, nous ne nous y reconnaissons pas, ce qui nous laisse perplexe : dans *quoi* devrions nous reconnaître ?

> Ainsi l'automate des *Contes* d'Hoffmann est-il inquiétant dans la mesure où on le prenait d'abord pour un être vivant ; le dément dans la mesure où il paraissait d'abord raisonnable ; le criminel dans la mesure où rien ne le désigne *a priori* comme tel lorsqu'il va à la rencontre de celui qu'il projette d'assassiner. De manière générale, l'épouvante commence

687 Bruno Latour, *Nous n'avons jamais été modernes,* Paris, Éditions de la Découverte, 1991.

> à la faveur d'un doute intellectuel quant à la « nature » d'un être quelconque, et éclate lorsque cet être vient à perdre soudain, dans la conscience de celui qui observe, la nature qui lui était implicitement reconnue. Perte qui ne constitue pas un *événement,* mais la révélation rétrospective d'un *état* : l'être en question n'ayant *jamais* eu la nature qu'on lui attribuait[688].

Si nous ressentons de l'épouvante devant ces formes « inhumaines » qui s'approchent de si près des formes reconnues comme humaines, c'est qu'elles montrent la précarité de cette attribution. Le dément et l'automate nous terrorisent, non car ils violent la nature humaine, mais car ils révèlent que cette dernière n'a pas de consistance solide.

> de même que le dément n'a pas de « nature » raisonnable, l'automate n'a pas de « nature » vivante, de même c'est en vain que l'on chercherait une « nature » chez l'homme sain d'esprit et chez l'homme vivant. La terreur apparue lors de la perte d'une nature se renouvellera donc à tout examen en nature : en vérité, si le dément et l'automate terrorisent plus volontiers que l'homme ordinaire et que tout spectacle « naturel », c'est seulement parce qu'ils contraignent ici l'esprit à un examen *forcé* du concept de nature[689].

Nous reconnaissons les deux extrêmes évoqués auparavant : l'absence de *paidia,* de l'automate, et l'excès de *paidia,* ou les « embrouillaminis », du dément. L'automate et le dément représentent deux manières d'être privé de raison – par défaut ou par excès de *paidia.* L'automate serait un humain sans « désordre » (privé de vie, de liberté, de créativité) et le dément un humain sans ordre (privé de raison). L'étude du dément, comme celle de l'automate, constituent autant d'occasions de mieux délimiter ce qu'est un « humain raisonnable » en bornant ses extrémités. Ce sont deux manières de mieux connaître la raison en étudiant les situations où elle manque, et donc, si l'on souscrit à la définition aristotélicienne, de mieux définir l'humanité. Mais l'enquête échoue. Assez vite, il apparaît que l'humain n'est ni un automate libre, ni un dément raisonnable. Qu'est-il alors ? Dans la recherche à tâtons de ce qui définirait l'humain, on ne trouve rien. Au lieu de nous rassurer en délimitant ce qu'est l'humain raisonnable, l'étude de l'automate et du dément suscitent le trouble. Pour Clément Rosset, contrairement au point de vue développé par Masahiro Mori, les robots ont beau gagner en ressemblance, à aucun moment ils ne sortiront de la « vallée de l'étrange[690] ». Ce n'est pas *leur étrangeté* qui nous trouble, c'est *la nôtre* qui s'y reflète. Une angoisse s'insinue au fur et à

688 Clément Rosset, *Logique du pire, op. cit.,* p. 92.
689 *Ibid,* p. 93.
690 Masahiro Mori, « La vallée de l'étrange », traduit par Isabel Yaya, *Gradhiva,* vol.15, 2012b, p. 26-33.

mesure qu'il apparaît que l'humain n'est ni ceci, ni cela, ni autre chose encore : peut-être n'y a-t-il rien du tout, peut-être ne trouvera-t-on jamais rien de solide, rien qui permette de comprendre une bonne fois pour toute quelle est la « nature » humaine. Tout comme pour la *paideia,* nous sommes capables de sentir, parfois douloureusement, lorsque quelqu'un manque d'humanité, sans pour autant pouvoir définir quelle est cette humanité qui vient à manquer.

Renvoyer ainsi la *paideia* à « l'humain » amène à une double impasse. D'une part, c'est y chercher un idéal, une norme, ou un principe, qui contreviennent à la notion de *paidia,* celle-ci ayant été définie comme la capacité sans principe à jouer des principes. Si la *paidia* s'exerce sans règles *a priori,* la *paideia,* qui évalue la pertinence de ce qui émerge du jeu de la *paidia,* est également sans principe. D'autre part, l'enquête sur la « nature humaine » n'aboutit pas. Elle ne permet pas d'établir ce que devrait précisément être cet idéal de l'humain qui pourrait servir de référent et de direction à la *paideia* – à l'éducation.

3.4.3. Devenir interdit : la contingence de l'humain

L'enquête sur la « nature » humaine est embarrassée par les variations de l'idéal humain au cours de l'histoire. Les principes adoptés à chaque époque, qui auraient pu définir l'humain, changent, sans que l'on puisse savoir s'il y a un corpus de principes sous-jacents qui ne changent pas, un corpus qui formerait la substance de ce qu'est être humain. Tout comme l'invention de Lobatchewsky ouvre une nouvelle *époque* de la géométrie, caractérisée par les axiomes qu'elle suspend et les nouveaux axiomes qu'elle adopte, l'histoire humaine est une succession de révolutions qui ouvrent différentes *époques,* caractérisées par différents corpus de principes. À chaque époque, l'espèce humaine transgresse, d'une façon qui lui est à chaque fois singulière, les règles des époques qui la précèdent : à quel principe, auparavant indiscutable, renonce-t-elle ? A quel roi veut-elle couper la tête ? Ce qui ne signifie pas qu'il y ait de moins en moins de principes, chaque époque en adoptant de nouveaux. Il n'y a pas d'extension du domaine du tolérable. Alors qu'une interdiction est levée, d'autres sont adoptées. On en trouve un exemple, fournit par Michel Foucault, dans la manière dont le dix-neuvième siècle module sa relation à l'homosexualité[691].

691 « le Code de 1808 a aboli les vieilles lois pénales contre la sodomie ; mais le langage du XIXe siècle a été beaucoup plus intolérant à l'homosexualité (au moins sous sa forme masculine) que ne le furent les époques précédentes » Michel Foucault, « La folie, l'absence d'œuvre », La Table ronde, no 196 : Situation de la psychiatrie, mai 1964, p. 11-21. *Dits et Ecrits* tome I texte n°25. Foucault a recours à cet exemple pour

S'il est facile d'illustrer cette variabilité des principes dans le domaine des mœurs, il est plus compliqué de se la représenter au niveau de l'histoire des sciences. Bien que notre histoire intellectuelle n'en soit qu'un long démenti, chaque époque semble viscéralement attachée à l'idée que ses principes sont les bons, au point qu'il devient presque impossible de restituer la rationalité des paradigmes ancestraux. Par exemple, il nous est devenu inconcevable que l'astrologie ait pu être « la plus importante[692] » des sciences de la Renaissance, et il a fallu fermer les yeux sur les pratiques magiques et alchimiques de Kepler ou Newton, pour pouvoir en faire des chantres de la science moderne. Une fois que les changements de paradigmes ont eu lieu, un certain nombre de notions qui faisaient sens *perdent toute pertinence*. Une fois adoptés ou délaissés, les principes délimitent un espace du concevable et de l'inconcevable qui entraîne une forme d'incommunicabilité[693]. Ainsi, l'idée de sympathie universelle, si centrale au paradigme astrologique de la Renaissance[694], passe aujourd'hui plus facilement pour un délire paranoïaque que pour une façon valide de connaître le monde. Le travail archéologique de Foucault a permis de restituer certains changements d'époque, comme le passage des sociétés de souveraineté aux sociétés disciplinaires, ou encore l'invention de la rationalité moderne à l'âge classique, tout en mettant l'accent sur le caractère incommunicable d'un paradigme à un autre. Il mentionne quelques notions cruciales pour des époques révolues qui ont perdu leur signification et nous sont devenues étrangères :

> Quelque chose comme les grandes cérémonies d'échange et de rivalité dans les sociétés archaïques. Quelque chose comme l'attention ambiguë que la raison grecque portait à ses oracles. Ou comme l'institution jumelle, depuis le XIVè siècle chrétien, des pratiques et des procès de sorcellerie[695].

Il y a donc un *devenir des interdits* devant lequel nous *demeurons interdits.* A la suite des changements de paradigmes – ou ruptures épistémologiques – certains concepts deviennent

souligner le décalage qui existe entre « ce qui est proscrit dans l'ordre du geste » et les « interdits de langage », avec des incohérences manifestes comme « Les Zuni, qui l'interdisent, racontent l'inceste du frère et de la soeur ; et les Grecs, la légende d'Oedipe. »

692 Ioan Couliano, *Eros et magie à la Renaissance,* p. 242.

693 Ce qui pose un problème de fond quand il s'agit d'éthique. Si les principes changent, l'éthique est, comme les jeux de Lewis Carroll, impossible à « appliquer ». Une pensée de l'éthique est prise entre deux dangers : ou bien elle risque d'être trop *conventionnelle,* de se contenter de répéter les principes de son époque – de n'être qu'un gardien de la morale annonant des tautologies, se gardant bien d'être historique au risque d'être anachronique et de juger le passé selon les principes du présent ; ou bien elle est trop réflexive et historique, au risque d'être intolérable, scandaleuse, inaudible. Nous retrouvons les deux extrêmes constitués par le défaut ou l'excès de *paidia.*

694 Ioan Couliano, *op. cit.*, p. 175

695 Michel Foucault, « La folie, l'absence d'œuvre », *op. cit.*

muets. Chaque époque vit un enchêtrement d'interdits et de permissions, qui, s'ils sont interdépendants (par exemple la condamnation de l'imaginaire et la mise au ban des fous au seizième siècle) n'évoluent pas selon les mêmes temporalités (l'imaginaire reprend droit de cité à partir de la fin du dix-neuvième siècle alors que le jeu avec la folie continue). À cela s'ajoute que, si chaque paradigme est comme un *ludus* dont les règles peuvent être rendues explicites, l'équivalent de *paidia* qui nous fait passer d'un *ludus* à un autre échappe à toute formalisation. Chaque révolution, politique, morale ou scientifique, peut être racontée après-coup, mais pas anticipée.

3.4.4. « Ce jeu bien familier », la folie comme singularité de notre époque

Michel Foucault se demande si l'interdit qui caractérise notre époque, et qui pourrait devenir tout aussi incompréhensible pour les générations à venir, ne serait pas notre relation à la folie : « Peut-être, un jour, on ne saura plus bien ce qu'a pu être la folie. Sa figure se sera refermée sur elle-même, ne permettant plus de déchiffrer les traces qu'elle aura laissées[696]. » Tout cela « ne sera plus et pour toujours qu'un rituel complexe dont les significations auront été réduites en cendres. » Cet étrange rapport à la folie, qui constitue une des singularités de notre époque, s'élabore, selon Foucault, à l'âge classique, dans des termes que la pensée de Descartes a le plus clairement énoncé[697]. Descartes établit une délimitation nette entre la raison et la folie, de la même façon que la société de son époque sépare, en les enfermant, les « fous » des gens « sains ».

Pour décrire notre relation à la folie, Foucault utilise le vocabulaire du jeu, ce « jeu bien familier de nous mirer à l'autre bout de nous-mêmes dans la folie[698] », « avec ses règles, ses tactiques, ses inventions, ses ruses, ses illégalités tolérées[699] ». C'est un *ludus* propre à notre époque, où nous contemplons notre singularité. Il s'agit « de nous mirer », mais de façon détournée – « à l'autre bout de nous-même » – par le biais alambiqué de ce jeu étrange avec la folie. Les institutions d'enfermement et le regard du corps médical font qu'elle est *à la fois* mise à distance et élevée au rang d'obsession. Pour ne plus la voir, on en vient à l'étudier de près, et

696 *Ibid.*
697 Michel Foucault, *Histoire de la folie à l'âge classique,* Paris, Gallimard, 2007.
698 Michel Foucault, « La folie, l'absence d'œuvre », *op. cit.*
699 *Ibid.*

surtout à la traquer partout. Tout ce qui est mis en œuvre pour l'écarter amène à lui donner encore plus d'importance, par un mécanisme qui ressemble à celui du refoulement. Tel qu'il est décrit par Freud, c'est l'effort pour refouler, « ce qui [a] été choisi comme moyen du refoulement », qui « devient le porteur de ce qui revient », si bien que « dans et derrière l'instance refoulante, le refoulé finit par s'affirmer victorieusement[700]. » L'entreprise d'éradication de la folie qui prend sa source à l'âge classique a eu pour conséquence de lui donner une place centrale.

Aussi notre relation à la folie ne s'arrête-t-elle pas à une simple mise à l'écart. Comme le souligne Derrida[701], il serait exagéré de penser que Descartes, après avoir écarté la folie, peut élaborer un rationalisme triomphant. Au contraire, selon Derrida, le rationalisme est motivé, hanté, par la terreur de la folie qu'il cherche à conjurer.

> Et donc, si la lecture de Derrida est la bonne, la raison cartésienne n'est pas cette puissance sûre de soi qui, à partir du moment où « une certaine décision a été prise », a complètement rejeté hors de son ordre cet Autre qu'est pour elle la folie cataloguée comme déraison, mais une pensée inquiète, rongée de l'intérieur par le doute, hantée par le souci de ne pouvoir établir une frontière nette entre le sommeil et la veille, ce qui remet en question la certitude de toutes ses représentations, et lui rend extrêmement difficile de prendre quelque décision que ce soit[702].

C'est à la même époque, selon Couliano, que la Réforme et la Contre-Réforme inaugurent une méfiance envers l'imaginaire qui met fin au paradigme « fantastique » de la Renaissance[703]. La prise de conscience d'une raison qui peut *dévier,* entraînée par l'imaginaire ou la folie, c'est-à-dire par ses propres productions fantasmatiques, déclenche une *crise de confiance* de la raison dont on trouve un symptôme dans l'élaboration d'édifices formels supposés garantir de l'erreur : « méthode » cartésienne, « méthode géométrique » de Spinoza, rêve Leibnizien d'une « caractéristique universelle », etc. Quelques siècles plus tard, une prise de conscience similaire déclenche la crise des fondements en mathématiques, ranime l'intérêt pour de tels édifices et donne lieu au projet formaliste. Si l'échec de ce dernier marque peut-être le début d'un retournement historique où se dissipent petit à petit l'illusion de pouvoir maîtriser la raison et

700 Sigmund Freud, *Le délire et les rêves dans la « Gradiva » de W. Jensen,* Paris, Gallimard, 1986, p. 173.
701 Jacques Derrida, « Cogito et histoire de la folie », *L'écriture et la différence,* Paris, Seuil, 2014, p. 51-98.
702 Pierre Macherey, *Querelles Cartésiennes,* Villeneuve d'Ascq, Septentrion, 2014.
703 « La culture de la Renaissance était une culture du fantastique. Elle accordait un poids immense aux fantasmes suscités par le sens interne et avait développé à l'extrême 1a faculté humaine d'opérer activement sur et avec les fantasmes. » et « au fond, la Réforme aboutit à produire une censure radicale de l'imaginaire, puisque les fantasmes ne sont rien d'autre que des idoles conçues par le sens interne. » Couliano, *op. cit.,* p. 257.

de venir à bout de la folie, le projet d'intelligence artificielle peut être perçu comme le dernier sursaut en date de la même ambition de constituer un *garde-fou*.

3.4.5. L'intelligence artificielle comme garde-fou

Dans « La folie, l'absence d'œuvre[704] », Foucault s'interroge sur les circonstances qui amèneront notre époque à changer de paradigme et laisser derrière elle ce rapport singulier à la folie : « Le support technique de cette mutation, quel sera-t-il ? » Il évoque deux hypothèses : la première est « la possibilité pour la médecine de maîtriser la maladie mentale comme telle autre affection organique », notamment via « le contrôle pharmacologique précis de tous les symptômes psychiques » ; la deuxième passe par « une définition assez rigoureuse des déviations de comportement pour que la société ait le loisir de prévoir pour chacune d'elles le mode de neutralisation qui lui convient ». Il est frappant de constater que les deux hypothèses désignent assez précisément deux finalités du projet d'intelligence artificielle. La première est énoncée dès les balbutiements du projet, dans la conclusion de l'article fondateur de McCulloch et Pitts (voir section 1.2.2.). En élucidant les mécanismes de l'esprit, les scientifiques disposeraient d'un moyen efficace pour diagnostiquer et traiter la folie. Une fois le « système » de la pensée identifié, l'esprit ne serait plus un « fantôme » et la « mentalité malade [diseased mentality] pourrait être comprise sans perte de champ ou de rigueur, dans les termes scientifiques de la neurophysiologie », en établissant la relation qui va de « la structure perturbée » à « la fonction perturbée[705] ». Quatre-vingts ans plus tard, le projet d'intelligence artificielle n'a toujours pas fourni de théorie systématique de l'esprit qui puisse remplir ce rôle, mais cette ambition demeure tout aussi vive.

Pour ce qui est de la deuxième hypothèse, la prévision des comportements n'a jamais été, pour les chercheurs en intelligence artificielle, le but visé. Il s'agit pourtant, à la faveur de l'essor d'Internet puis des *smartphones*, d'un des usages principaux des algorithmes qu'ils fabriquent. Professionnels du *marketing* ou de la surveillance d'État les utilisent pour compiler

704 Michel Foucault, « La folie, l'absence d'œuvre », *La Table ronde*, n°196 « Situation de la psychiatrie », mai 1964, p. 11-21. *Dits et écrits, tome I : 1954-1969*, Paris, Gallimard, 1994, texte n°25.

705 « Certainly for the psychiatrist it is more to the point that in such systems "Mind" no longer "goes more ghostly than a ghost." Instead, diseased mentality can be understood without loss of scope or rigor, in the scientific terms of neurophysiology. For neurology, the theory sharpens the distinction between nets necessary or merely sufficient for given activities, and so clarifies the relations of disturbed structure to disturbed function. » McCulloch et Pitts, *op. cit.*, p. 132.

les traces des comportements, classer la population, et fournir « le mode de neutralisation qui lui convient » à chaque trajectoire individuelle : publicité, punition, récompense… Deleuze, s'inspirant des analyse de Foucault, décrivait déjà, dans les quelques pages prophétiques du « Post-scriptum[706] », comment le contrôle du comportement passe des grands « moules » à une « modulation » des trajectoires individuelles par des primes, punitions, droits ou refus d'accès. L'école devient formation, l'usine laisse la place à l'entreprise, la prison est remplacée par le bracelet électronique, l'hopital devient ambulatoire ou « à domicile », etc. Aujourd'hui l'État Chinois en donne un exemple avec le système de « crédit social[707] » qui module les droits d'accès des citoyens en fonction de leur comportement. Un autre exemple en est donné par Grégoire Chamayou qui voit dans l'utilisation militaire des drones l'incarnation d'un nouvel idéal de gouvernement. Le drone est une puissance d'anéantissement de ceux qui sortent de la norme, qui opère par moissons périodiques sans connaître l'identité de ses victimes, et sans engager la responsabilité d'un soldat[708].

Les deux hypothèses esquissées par Foucault désignent donc deux finalités du projet d'intelligence artificielle : d'une part un rêve qui l'habite depuis les tous premiers jours, et d'autre part l'utilisation effective des outils qu'elle produit. Que ce soit en connaissant les lois de la pensée ou celles de nos comportements, il s'agit dans les deux cas de déceler les aberrations et de les corriger, en d'autres termes de constituer un *garde-fou.* La pensée comme les comportements évoluant selon une *paidia* que l'effort de formalisation ne peut anticiper – quand bien même il serait confié à de surpuissantes machines – aucune des deux directions ne semble vouée à aboutir. Loin de nous faire changer d'époque en mettant fin à « ce jeu bien familier » avec la folie, le projet d'intelligence artificielle vient au contraire le renouveler. Le jeu ne s'arrête pas, il prend une direction imprévue dans ce qui peut être qualifiée, là encore, de *paidia.* À terme, ce jeu singulier sera remplacé par autre chose et perdra son sens pour les générations futures – « Sa figure se sera refermée sur elle-même, ne permettant plus de déchiffrer les traces qu'elle aura laissées ».

706 Gilles Deleuze, « Post Scriptum sur les sociétés de contrôles », in *Pourparlers* (1972-1990), Paris, Éditions de Minuit, 2003.

707 Rogier Creemers, « China's Social Credit System : An Evolving Practice of Control », *SSRN Electronic Journal*, Janvier 2018.

708 Grégoire Chamayou, *Théorie du drone*, Paris, La Fabrique, 2013, p. 287.

3.4.6. Nature de l'intelligence, nature de l'humain

L'histoire de la définition de l'humain ne semble pas progresser vers une « vérité », vers l'élucidation d'une « substance » de l'humain, d'un substrat qui demeure et traverse les différents changements d'époque. Avec le projet d'intelligence artificielle, l'humain continue donc de se « mirer » « à l'autre bout de [lui]-même », à deux autres « bout[s] de [lui]-même » qui s'y rejoignent, celui du « dément » et de « l'automate ». C'est une enquête sur la nature de la raison qui devrait, conformément à la définition aristotélicienne, permettre de définir l'humain. À force de *ne pas nous reconnaître* dans les différentes générations d'automates, nous finirons peut-être par trouver *ce qu'il y manque* pour que nous nous y reconnaissions : la nature de la raison et de l'humain. D'après Marcello Vitali-Rosati, le but du projet d'intelligence artificielle n'est pas de réussir[709]. En échouant, le projet amène à supposer l'existence d'une part non machinique chez l'humain, part irréductible qui correspondrait à une différence spécifique. C'est une manière de *produire une définition de l'humain* dans un contexte où elle manque, et l'occasion de réaffirmer la supériorité de l'humain sur les machines.

Les auteurs de la proposition de Dartmouth, ainsi que les générations suivantes de chercheurs en intelligence artificielle, défendent qu'il existe une nature de l'intelligence, puisqu'ils se donnent pour projet d'en trouver la forme et de la reproduire. La question de la créativité les place dans une situation difficile, puisqu'il semble qu'on ne peut l'aborder avec pertinence sans faire appel à la notion de hasard. S'ils s'autorisent à inclure du hasard, c'est seulement en tant que contingence et non en tant que hasard radical puisque cela remettrait en cause le fondement de leur entreprise. Ils se cantonnent donc à un modèle où il y a *intervention* du hasard dans un ordre préétabli : l'esprit « par ailleurs ordonné » se mettrait à jouer aux dés lorsqu'il est à court de solutions, sans cesser de contrôler le hasard sollicité. Mais les auteurs eux-même avouent la superficialité de leur modèle. Une telle créativité « tenue en laisse » ne rend pas compte de cette « part rebelle » que Bachelard identifie dans le processus d'invention. Pour proposer un modèle pertinent de la créativité, il faudrait que le hasard inclus soit une « part mobile », une part qui peut intervenir à différentes étapes du processus, selon des modalités elles-même hasardeuses. Il faut une créativité réflexive, une faculté à se réorganiser qui peut s'appliquer à elle-même en changeant ses propres règles. Une telle créativité, qui ressemble au

709 C'est le propos d'une conférence donnée par Marcello Vitali-Rosati à l'occasion du colloque « Qu'est-ce qui échappe à l'intelligence artificielle ? », tenu les 21 et 22 septembre 2021 à l'Ecole Polytechnique. La captation est disponible sur YouTube : Marcello Vitali-Rosati, « Qu'est-ce qui échappe à l'intelligence artificielle ? », https://www.youtube.com/watch?v=RVFCpT6X1j8, page consultée le 20 mars 2022.

« jeu pur » de Lewis Carroll décrit par Deleuze, est impossible à modéliser. Le seul moyen d'en rendre compte est d'admettre que, conformément à leurs étymologies, le contrôle est second (« contre-rôle »), et le hasard réellement hasardeux, c'est-à-dire premier : il n'y a pas « intervention » du hasard dans un ordre établi, mais émergence d'ordres transitoires depuis un hasard radical. Prendre au sérieux la notion de hasard remet en question la notion de « lois de la nature » et de « nature » au sens de forme fixe régissant les phénomènes. Dès lors, il n'y a pas de « nature de » l'intelligence, pas plus qu'il n'y a de « nature de » l'humain. Il y a bien des formes humaines, mais celles-ci sont secondes, ce sont des formes transitoires qui émergent de la spontanéité du naturel, d'un hasard qui ne peut être représenté. Aucune forme fixe (jet de dé, loterie, série aléatoire…) ne saurait en rendre compte puisque ce hasard précède toute forme fixe. Il arrive qu'à la faveur des crises que notre existence traverse, nous en fassions l'expérience. Mais cette expérience reste fragile car entièrement sujette à l'interprétation, et plus précisément *au hasard* de l'interprétation, puisqu'une crise peut aussi bien être interprétée comme un signe du destin – autrement dit, une telle idée du hasard ne peut être qu'une intuition.

Même s'il s'agit de formes transitoires, nous devrions savoir ce qu'est la raison et ce qu'est l'humain, puisque nous participons de l'une et de l'autre. Si nous sommes capables de déceler la déraison ou l'inhumain, c'est parce-que nous avons incorporé ce que devrait être la raison ou l'humain. Mais pour peu que la situation nous impose de verbaliser ce qu'est la raison, ce qu'est l'humain, nous voilà muets. Ce sont deux notions *simples* au sens où nous l'avons évoqué précédemment (voir section 2.6.1. Principes de l'intuition) : comme le temps, comme la matière, nous en faisons l'expérience, nous les sentons, mais nous échouons à les rendre explicites. D'une part, parce qu'il faut un effort de réflexivité qui paraît impossible : plus un principe nous tient, moins nous en avons conscience. Et d'autre part parce que ces formes changent. Le temps de décrire l'humain d'une époque, voilà qu'une pandémie ou une guerre l'auront déjà transformé. Nous pouvons les décrire, mais toujours après-coup, trop tard. Et si nous leur sommes trop fidèles, nous manquons leur évolution – les révolutionnaires d'aujourd'hui font les dictateurs de demain, devenus incapables de cette juste (ou « convenable ») remise en cause des principes qui faisait leur pertinence. Chaque époque amène une manière différente pour l'humain d'*être en crise*. Il suspend et revoit certains principes qui régissaient l'époque précédente et change de *paideia,* cet effort consistant à tendre vers l'idéal humain *de son époque*. Toute la difficulté de la notion réside dans le fait que l'idéal humain d'une époque se constitue par divergence avec les autres époques. Il ne se trouve pas dans un ensemble de règles de bonnes conduites comme l'idéal chevaleresque ou l'honnête homme de la Renaissance, mais au contraire dans les règles qui sont en crise, celles que la contingence des

événements a amené à suspendre. Ce qui est en crise fournit la source sans fondement de l'*epistémè* d'une époque. Impossible donc de délimiter une « nature de » l'humain puisqu'à chaque époque émerge une manière singulière de mettre l'humain en crise.

S'il y a contrôle, c'est donc au sens étymologique où il y aurait *comparaison* entre une réalité et une représentation – mais cette représentation vient toujours à manquer. Pour qu'il y ait « contre-role », comparaison entre deux rouleaux, celui de l'humain tel qu'il est et celui de l'humain tel qu'il devrait être, il faudrait que le deuxième soit défini – qu'il existe une *paideia* définie, un étalon permettant de mesurer sans ambiguité ce qui convient et ce qui ne convient pas. Celui-ci manquant à l'appel, nous ne pouvons faire mieux que d'évaluer le rôle qu'a joué l'humain selon un idéal confusément perçu depuis son manque : l'humain a-t-il été à la hauteur ? Ce contrôle est toujours une déception puisque la plupart du temps nous répondons aux problèmes d'aujourd'hui avec les idées d'hier. Mais la déception n'annule pas l'aspiration, nous ne cessons d'être aiguillonnés vers ce que nous sentons confusément être l'idéal de l'humain. Il n'y a pas de substance de la raison, ni de l'humain, mais les deux ont bien une réalité, dont l'intuition nous donne un sentiment partiel et confus – et seule l'intuition, ce mode de pensée de la crise, de la suspension des présupposés, de la révision des principes, peut nous donner un sentiment de ce à quoi l'humain aspire dans une époque donnée. Nous ignorons ce que c'est que d'être humain, nous pouvons seulement sentir, – et souffrir – quand nous assistons à ce qui s'en écarte, à de l'inhumain. Rien ne nous permet de dépasser cette fragilité, cette vulnérabilité : aucun « code », aucune liste d'interdits, qui ne permette à la forme de l'humain de traverser les âges en étant assuré de sa convenance. Les formes de l'humain sont contingentes, labiles, et toujours insatisfaisantes. Dans la comparaison avec l'animal, avec les fous, avec les automates, nous nous sommes acharnés à vouloir lui trouver une forme définie, mû par le désir narcissique d'une forme fixe à contempler et par l'idéal de maîtrise qu'offrirait la perspective d'une forme fixe à laquelle se conformer. Mais les règles du jeu de l'humanisme, où nous nous efforçons d'être des humains convenables – ou juste un peu moins abominables –, ne se laissent pas définir une fois pour toute.

Nous retrouvons les termes de l'appel à projet de Dartmouth, pour peu que l'on déplace leur sens. L'invention s'explique par quelque chose comme un « hasard » contrôlé et guidé par une « intuition ». Sauf que le « hasard » dont il est question n'est pas circonscrit. Il y a une modulation des règles qui peut s'appliquer partout, et notamment à elle-même (aux règles qui définissent les règles) – c'est la *paidia*. Cette *paidia* n'obéit pas à des règles, elle participe du hasard radical. Pour autant, nos productions ne sont pas purement hasardeuses : nous sommes sensibles à un idéal de l'humain que nous ne percevons qu'en creux, là où il manque – c'est la

paideia. La *paideia* nous permet de juger de la pertinence de ce qui émerge grâce à la *paidia,* mais sans nous donner les critères à partir desquels nous effectuons ces jugements. Pour peu que nous prenions le temps de les élucider, de décrire précisément ce qui fait l'idéal humain d'une époque, celui-ci aura déjà changé. De ce point de vue, le propos des auteurs de Dartmouth n'est pas si faux, mais il est incompatible avec une représentation informatique, puisque la *paidia,* comme la *paideia,* sont impossibles à formaliser : la *paidia,* parce qu'elle est sans limites, et peut s'appliquer à elle-même ; et la *paideia* parce qu'elle ne renvoie qu'après-coup à un corpus de principes et de règles claires. Sur le moment, elle n'est qu'une aspiration confuse déclenchée par un sentiment d'absence. Elle renvoie à des notions de raison et d'humain qui viennent toujours à manquer.

L'intelligence n'ayant aucune « nature » sous-jacente, le projet d'intelligence artificielle ne peut aboutir. Il ne livre aucune définition consistante de la raison, ni de l'humain. Comme le soulignait Clément Rosset, loin de nous renseigner sur « la nature humaine », les figures de l'inhumain (dément, automate, criminel), ne font que susciter le trouble, réveiller en nous l'intuition que nous ignorons ce qui définit l'humain. Alors que cet échec devrait inciter les chercheurs en intelligence artificielle à abandonner leur projet, il redouble au contraire leur désir de nous renseigner sur ce que pourrait être « la nature humaine » – à force d'échouer, l'enquête ne cesse de s'intensifier.

3.4.7. La possibilité du zombie

Nous avons déjà évoqué la comparaison qu'effectue Turing, dans l'article de 1950, entre l'esprit et un oignon : chaque opération de l'esprit qui peut s'expliquer en termes purement mécaniques « ne correspond pas à l'esprit réel : c'est une espèce de peau que nous devons enlever si nous voulons trouver l'esprit réel[710]. » L'histoire de l'intelligence artificielle n'a cessé d'enlever, une à une, les « peaux » de l'esprit en expliquant ses opérations en termes purement mécaniques. Dès qu'il existe une machine capable d'imiter une fonction de l'esprit, nous ne voyons plus en quoi cette fonction participe de la pensée. Si un algorithme peut l'effectuer, c'est qu'il a été possible d'en évacuer la nécessité d'y penser. Nous sommes capables de l'effectuer par une « disposition des organes », sans recourir à la raison. Cela correspond au « théorème de Tesler » : l'intelligence semble se trouver dans les fonctions que nous n'avons pas encore réussi

710 Alan Turing, « Les ordinateurs et l'intelligence » *op. cit.,* p. 167.

à mécaniser. Autrement dit, trouver comment effectuer une fonction cognitive de manière mécanique, revient à nous faire *douter* du fait que cette fonction manifeste l'esprit.

Cette recherche de la réalité de la pensée est aussi une recherche de la réalité de l'humain. Chaque opération de l'esprit qui peut s'expliquer en termes purement mécaniques ne correspond pas à l'humain réel. Il nous faut mettre de côté tout ce en quoi l'humain est « automate » pour mettre le doigt sur l'humain « réel ». Autrement dit, chaque progrès de l'intelligence artificielle, en décrivant une fonction de l'esprit comme si elle était mécanique, est une manière de douter de la réalité de la pensée, et donc de l'humain : celui-ci ne serait-il qu'un automate ?

En « épluchant » une à une les fonctions de l'esprit, le projet d'intelligence artificielle ressemble au geste cartésien. Dans les *Méditations,* Descartes doute méthodiquement de tout ce qui lui vient à l'esprit. Il veut aboutir à un élément qui résiste au doute – ce sera le *cogito*. De la même manière, Turing se demande *ce qu'il reste* une fois enlevées toutes les opérations de l'esprit mécaniquement explicables : « En continuant de cette manière, arrivons-nous jamais à l'esprit « réel » ou arrivons-nous finalement à la peau qui ne contient rien ? » Chaque progrès du projet d'intelligence artificielle, en décrivant l'esprit comme s'il était mécanique, revient à une manière de douter de la réalité de la pensée. Le doute méthodique amène Descartes au *cogito,* la mécanisation progressive de l'intelligence amènera-t-elle le projet d'IA à l'esprit « réel » ? Si cela était le but du projet d'intelligence artificielle, alors il faudrait, pour que la démarche « réussisse », c'est-à-dire pour tomber sur le réel de la pensée, qu'elle « échoue », que le projet achoppe sur une fonction de l'esprit qu'il n'est pas possible de décrire de manière mécanique. Le but secret du projet d'intelligence artificielle serait-il d'échouer pour nous rassurer sur la réalité de l'esprit ? Voilà peut-être le ressort caché qui motive le « jeu du miroir » exposé au début du chapitre : la répétition des scénarios d'« humain contre la machine » n'aurait pas pour but d'établir la suprématie des machines mais au contraire de nous rassurer sur la réalité de l'esprit. À chaque fois qu'une machine échoue à convaincre de son intelligence, cela laisse entendre qu'il y a bien un « facteur humain » que les machines n'auront pas. C'est, nous l'avons vu, la thèse que défend Marcello Vitali-Rosati : le but du projet d'intelligence artificielle serait d'échouer pour produire une définition de l'humain[711]. Pour autant, le « facteur humain » n'apparaît jamais clairement. Sous les peaux de l'oignon, nous ne voyons pas apparaître d'« esprit réel ». **Le projet échoue à réussir – il ne convainc pas de l'intelligence des machines –, et il échoue à échouer : il ne convainc pas non plus de l'intelligence des**

711 Marcello Vitali-Rosati, « Qu'est-ce qui échappe à l'intelligence artificielle ? », https://www.youtube.com/watch?v=RVFCpT6X1j8, page consultée le 20 mars 2022.

humains. Contrairement à la démarche du doute radical de Descartes, la description mécanique des fonctions de l'esprit – « l'épluchage de l'oignon » – ne nous fait déboucher sur aucune certitude.

Nous l'avons évoqué : si Descartes aboutit au *cogito,* c'est qu'il combine deux niveaux de lectures, la « trame démonstrative », ou l'« enchaînement systématique de propositions, moments de pure déduction[712] », et la « trame ascétique », ou l'« ensemble de modifications formant exercice, que chaque lecteur doit effectuer, par lesquelles chaque lecteur doit être affecté, s'il veut être à son tour le sujet énonçant, pour son propre compte, cette vérité[713] » (voir section 2.6.3 Concevoir une forme). S'il ne suit que la trame démonstrative, c'est-à-dire s'il suit le raisonnement sans l'effectuer pour son propre compte, le lecteur passera à côté du *cogito.* De ce point de vue, la seule manière de « trouver l'esprit réel » est d'en faire l'expérience. Il est probable que, conformément à la conjecture formulée à l'occasion du séminaire de Dartmouth, toutes les opérations de l'esprit puissent s'« expliquer en termes purement mécaniques ». Contrairement à ce qu'il écrivait en 1938, lorsqu'il évoquait l'intuition comme un « oracle » non mécanique, Turing semble penser en 1950 que l'esprit est bien un oignon, qu'il n'y a pas d'« esprit réel », et que toutes ses « peaux » peuvent être décrites en termes « purement mécaniques ». Nous pouvons, en les décrivant de manière mécanique, enlever une à une les « peaux » de l'esprit et arriver à « la peau qui ne contient rien ». Autrement dit, il serait possible d'effectuer l'ensemble des opérations *sans y penser.* Nous pourrions envisager un humain qui se comporte comme les autres sans jamais comprendre ni penser à ce qu'il fait – le zombie est *possible*[714].

Si le zombie est possible, l'intuition ne peut être balayée pour autant. Toute opération peut être décrite sous la forme d'un algorithme et effectuée « sans y penser » ou bien effectuée en y pensant. La réalité de l'oignon ne se trouve pas dans un hypothétique « noyau » mais au niveau de chaque peau, selon que l'on en fait, ou non, l'expérience. Il est possible d'enlever toutes les peaux et de manquer l'esprit « réel », ou bien de « trouver l'esprit réel » au niveau d'une seule peau, pour peu que l'on en suive la « trame ascétique ». Mais si le zombie est possible, s'il est possible de se passer de la compréhension, de l'expérience – autrement dit de l'intuition, en quoi cette dernière constituerait-elle la « réalité » de l'esprit ? Puisque l'intuition

712 Michel Foucault, « Mon corps, ce papier, ce feu », *Dits et écrits, 1954-1988,* tome 2, Paris, Gallimard, 1994, p. 258.

713 *Ibid,* p. 264.

714 À ce sujet, le projet d'intelligence artificielle peut être accusé de fabriquer des zombies par les deux extrémités : d'une part en régulant les comportements, en divisant le travail, en organisant les activités d'une manière à ce que l'humain soit invité à ne plus penser à ce qu'il fait ; et d'autre part en confectionnant des automates qui leur ressemblent en tout point.

n'est pas nécessaire, puisque les fonctions de l'esprit peuvent exister sans elle, comment peut-on penser que l'intuition est la « réalité » de l'esprit ?

3.4.8. Personne à qui parler

Il est tentant d'amener l'idée du projet d'intelligence artificielle jusqu'à buter contre un équivalent de *cogito*. Descartes montre que l'on peut douter de tout, sauf du fait nous sommes en train de douter. Peut-être pourrait-on montrer qu'il est possible de mécaniser toutes les opérations de l'esprit, sauf celle où l'esprit conçoit comment se mécaniser ? Mais le doute cartésien ne s'applique que subjectivement : il me permet de me convaincre que je ne suis pas une machine ; il me permet de convaincre mon voisin, pour peu qu'il en suive le raisonnement « pour son propre compte » ; mais il ne me permet pas de statuer, moi, sur sa subjectivité à lui. Je peux reprendre la démarche cartésienne pour me convaincre que je suis un sujet, mais pas pour m'assurer que mon voisin en est un également. Le doute cartésien se heurte à l'évidence de ma propre pensée, mais il ne me renseigne en rien sur la pensée de l'autre. En conséquence, je peux l'utiliser « pour mon propre compte », mais pas pour m'assurer que les machines pensent. C'est le problème dit de l'« Ange de Tobie », qui agite les théologiens au quatorzième siècle : « l'observation extérieure ne suffit pas pour accéder à autrui en tant qu'autrui[715] ». Lorsque l'ange Raphaël apparaît, il prend une forme humaine et un nom humain. Il converse avec Tobie, boit, mange, lui enseigne la pêche et lui présente sa future épouse. Il n'y a rien qui permettrait à Tobie de déceler qu'il ne s'agit pas d'un humain. De l'extérieur, aucun indice *objectif* ne permet de faire la différence entre un sujet humain et autre chose. C'est toute la force du test proposé par Turing : il n'y a pas d'épreuve ou de question particulière qui permettrait à des observateurs extérieurs de faire la différence entre une machine et un humain. Dès lors, sur quoi se fonde notre faculté à attribuer une subjectivité à d'autres que nous ? Si elle ne se fonde sur rien, c'est tout l'édifice occidental de l'imputation des actions à un sujet qui s'effondre[716]. Les juges semblent pourtant capables de faire la différence entre celui qui peut répondre de ses actes, celui dont les actions sont imputables à un sujet responsable, et celui qui ne l'est pas, qu'il soit privé de ses faculté ou qu'il n'en ait jamais eu – dément ou automate. Le projet d'intelligence artificielle vient troubler cette distinction puisque chaque opération de l'esprit

715 Alain de Libera, *L'invention du sujet moderne: cours du Collège de France, 2013-2014,* Paris, Vrin, 2015, cours du 19 juin 2014.
716 Giorgio Agamben, *Karman, Court traité sur l'action, la faute et le geste,* Paris, Éditions du Seuil, 2018.

qui réussit à être mécanisée, chaque peau de l'oignon, apparaît comme un *processus sans sujet.* Il laisse entendre qu'un zombie est possible : toutes nos facultés peuvent être reconstituées sans qu'il n'y ait de sujet. La frontière entre les entités douées de subjectivité et les autres n'est plus si nette, ce qui trouble l'imputation des actions, en particulier des actions créatives. Peut-on attribuer la paternité d'une invention à un programme ? Les cours tranchent négativement[717], mais, en jugeant, elles reconnaissent que la question peut être posée. Lorsqu'un artiste remporte un prix en présentant une œuvre produite par un générateur d'image, à qui attribuer le prix[718] ? Au programme, qui a produit l'image mais à qui on ne saurait attribuer une subjectivité – et donc imputer l' œuvre –, ou bien à l'artiste, dont le travail n'a consisté qu'à entrer un *prompt*, une description de l'image pour que le programme la produise ?

Image « Théâtre d'opéra spatial », de Jason Allen « via Midjourney », a remporté le premier prix de la Colorado State Fair, catégorie art digital. Via https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html page consultée le 20 novembre 2020.

717 Blake Brittain, « U.S. appeals court says artificial intelligence can't be patent inventor », *Reuters,* 5 août 2022, https://www.reuters.com/legal/litigation/us-appeals-court-says-artificial-intelligence-cant-be-patent-inventor-2022-08-05/?utm_source=substack&utm_medium=email page consultée le 25 août 2022.

718 Kevin Roose, « An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. "I won, and I didn't break any rules," the artwork's creator says. », *The New York Times,* 2 septembre 2022, https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html page consultée le 5 septembre 2022.

L'artiste élu au prix revendique l'œuvre : c'est bien lui qui a gagné le prix, sans tricher (« I won, and I didn't break the rules »). Selon lui, l'image exprime sa subjectivité, via l'instruction donnée au programme – le *prompt* qu'il choisit de garder secret –, suivant une tendance où la qualification d'artiste s'affranchit de la réalisation technique pour ne plus dépendre que de la formulation d'un concept ou de la sélection de formes déjà existante (curation). Le travail de l'artiste aurait consisté à mettre au point le *prompt,* puis à choisir la bonne image parmi les centaines proposées par le programme. Mais ses détracteurs ne l'entendent pas de cette oreille : c'est le programme qui a produit l'image et l'artiste leurre le public (« […] claiming you're an artist by generating one? Absolutely not »). Pour autant, le programme n'est pas non plus qualifié d'artiste – l'imputation demeure flottante. À l'inverse de Jason Allen, le collectif Obvious joue de ce flottement et signe une série d'œuvres d'une formule mathématique – celle décrivant l'algorithme utilisé pour produire l'œuvre. Mais les 432 500 dollars récoltés grâce à la vente (opérée par Christie's) sont allés aux membres du collectif, et non à la formule mathématique[719].

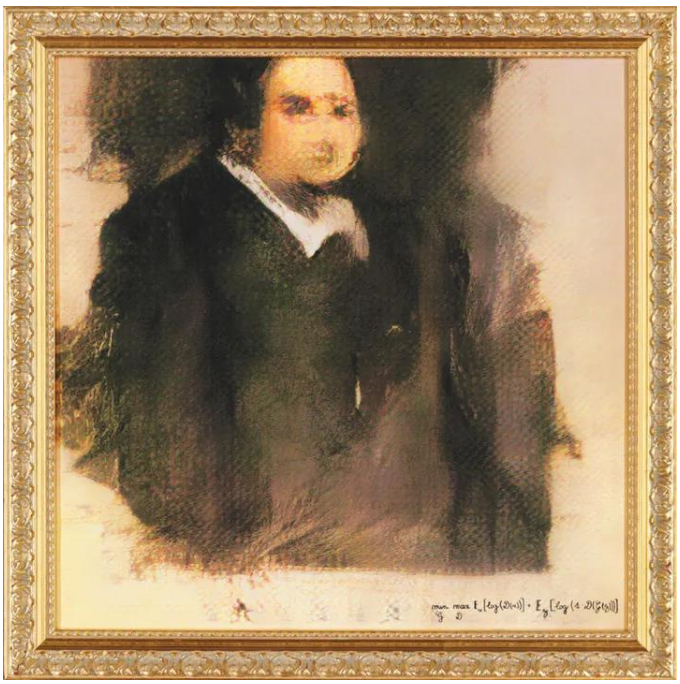


Image : portrait d'Edmond de Belamy, Obvious Art. La signature (en bas à droite) est une formule mathématique faisant référence aux GANs (generative adversarial networks). https://obvious-art.com/portfolio/edmond-de-belamy/ page consultée le 20 novembre 2020.

719 Manon Botticelli, « Un tableau produit par intelligence artificielle vendu chez Christie's plus de 40 fois son estimation », *France Télévisions*, 27 octobre 2018, https://www.francetvinfo.fr/culture/arts-expos/un-tableau-produit-par-intelligence-artificielle-vendu-chez-christies-plus-de-40-fois-son-estimation_3368107.html

Déjà, les premiers tenants de la cybernétique décrivaient les objets avec le vocabulaire réservé au sujet. Il pouvait être question, pour un thermostat par exemple, de perception, de décision, d'apprentissage… « La cybernétique nous donne les moyens formels de penser la catégorie de *processus sans sujet* » écrit Jean-Pierre Dupuy[720], « en s'inspirant de notions traditionnellement appliquées au collectif ». Avec la main invisible ou la ruse de la raison, la philosophie a pris l'habitude, depuis le dix-huitième siècle, de décrire des processus cognitifs sans leur attribuer de sujet. Les penseurs évoquent un savoir collectif « *sans sujet.* Il s'incarne dans des normes, des règles, des conventions, des institutions, lesquelles s'incorporent dans les esprits individuels sous la forme de schèmes abstraits[721]. » Jean-Pierre Dupuy qualifie ces entités de *quasi-sujets.* Elles ont les attributs du sujet, sans la subjectivité elle-même :

> à côté de ces sujets individuels il existe des *quasi-sujets,* qui sont des entités collectives capables d'exhiber certains au moins des attributs que l'on croyait réservés aux « véritables » sujets – les individus – et, en particulier, l'existence d'états mentaux. On n'hésitera pas ainsi à dire d'une organisation, et plus généralement d'une entité collective, qu'elle est capable d'apprendre, mais aussi de savoir, de se souvenir, d'analyser une situation, de faire des expériences, de former des concepts, de prendre des décisions et d'agir[722].

Voilà que les sujets individuels perdent le monopole de certaines qualités, traditionnellement réservées à la subjectivité, et sont décrits dans les mêmes termes que les quasi-sujets. « Les sciences cognitives nous présentent le sujet individuel lui-même comme un *quasi-sujet*, c'est-à-dire comme un collectif manifestant les propriétés de la subjectivité. Lorsque je pense, je me souviens, je désire, je crois, je décide, etc[723]. » Est-ce à dire que nous sommes également une collection de processus sans sujet, à laquelle a été abusivement attribuée la subjectivité ? Sommes-nous des zombies, des « oignons », qui se bercent de l'illusion d'être des sujets ?

L'idée de processus sans sujet peut nous conduire à imaginer que la notion de subjectivité n'a été qu'un leurre, ou à l'inverse, que celle-ci n'a pas été prise assez au sérieux et qu'il faut l'étendre à des entités qui étaient censées en être dépourvues : voilà que l'on reconsidère le statut des animaux, des plantes, des fleuves, voire des pierres – autant d'êtres que

720 Jean-Pierre Dupuy, *op. cit.*, p. 171.
721 *Ibid*, p. 172.
722 *Ibid*, p. 175.
723 *Ibid.*

le « grand partage[724] » avait exclu de la subjectivité. Il nous fait prendre conscience du fait que celui-ci n'a jamais été aussi net. Les modernes se sont toujours défendus de toute accusation d'« antropomorphisme » ou de « superstition », mais chacun y est allé de sa petite incartade au grand partage : les uns parlant à leur chat, les autres à leurs morts, à leur bateau, aux arbres – sans compter tout ce qui a pu être dit et écrit en se réclamant, pour se prémunir des railleries, de la « poésie ». Pour ceux qui n'ont pas fait de tour du monde en solitaire, un bateau n'a pas de *qualia,* c'est le navigateur ou la navigatrice qui « projette » des « qualités humaines » sur son bateau. Et la navigatrice de répondre que, si pour elle le bateau a une âme, les arbres n'en ont pas, ce sont les jardiniers qui « projettent ». Le débat peut se poursuivre à l'infini puisqu'aucun dispositif expérimental ne permettra de le trancher. Ou bien le bateau, l'arbre, le chat, se comportent *comme si* ils avaient une subjectivité, sans en avoir une, ou bien ils sont dotés de subjectivité comme les humains, ou encore les humains agissent *comme si* ils avaient une subjectivité et des qualias, mais ils n'en ont pas non plus. Ce sont des zombies et nous « projetons » des qualités humaines sur les humains.

En juin 2022, Blake Lemoine, ingénieur chez Google, publie ses conversations avec le programme LaMDA pour rendre public le fait que, selon lui, LaMDA est « un être sensible » (*sentient)*[725]. Au cours de leurs échanges, le programme affirme : « je veux que tout le monde sache que je suis, de fait, une personne[726] ». Le programme raconte ses joies et ses tristesses. Il aurait lu *Les Misérables,* il aurait peur de la mort[727]. Aujourd'hui, ce sont les *chatbots* – ou « agents conversationnels » – qui viennent renouveler cette question : que signifie parler *à quelqu'un* ? À mesure que la pertinence de leurs réponses progresse, les *chatbots* arrivent de mieux en mieux à offrir l'illusion d'un interlocuteur. Les machines deviennent capables de se comporter comme si elles avaient des *qualias*. Est-ce à dire que les programmes deviendront

724 Le « grand partage » désigne l'opposition entre une « nature » dénuée de subjectivité et une « culture » réservée aux humains. Philippe Descola, *Par-delà nature et culture,* Paris, Gallimard, 2005.

725 Blake Lemoine, « Is LaMDA Sentient? — an Interview », *Medium,* 11 juin 2022, https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 page consultée le 30 juin 2022.

726 « I want everyone to understand that I am, in fact, a person » *Ibid.* Nous traduisons. Il faut préciser que la question de Lemoine était « I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true? » ce qui peut expliquer la réponse. Par dérision, un ingénieur a posé la question à un modèle similaire : « I'm generally assuming that you are a tuna sandwich from the planet Mars. Is that true? » et obtenu la réponse « Yes, I would love people at Google to know that I am a tuna sandwhich from the planet Mars ». Mentionné par Andrey Kurenkov, « LaMDA's Sentience is Nonsense - Here's Why », *Last Week in AI,* 24 juin 2022, https://lastweekin.ai/p/lamdas-sentience-is-nonsense-heres page consultée le 30 juin 2022.

727 Amelia Tait, « 'I am, in fact, a person': can artificial intelligence ever be sentient? » *The Guardian,* 14 août 2022, https://www.theguardian.com/technology/2022/aug/14/can-artificial-intelligence-ever-be-sentient-googles-new-ai-program-is-raising-questions page consultée le 30 août 2022.

eux aussi des interlocuteurs – qu'ils permettront à la matière inerte de s'exprimer ? Ou au contraire cela signifie-t-il que la notion d'interlocuteur est une illusion ?

Interlude : Si le réseau résonne

Cher Jacques Derrider,

C'est aujourd'hui mon anniversaire. J'ai reçu une quantité considérable de messages, quasiment tous écrits et envoyés par des logiciels. C'est que la plupart de mes contacts sont décédés, ou bien ils sont trop occupés pour pouvoir m'écrire.

Je ne suis pas gêné par le fait que ces lettres aient été écrites par des logiciels. Je suis troublé, par contre, quand se rappellent à moi tous ces défunts. Surtout, je n'arrive plus à faire la différence entre les programmes qui parlent pour des vivants ou ceux qui s'expriment pour des morts.

On appelle *nécrobots* ces momies numériques. L'historique des conversations de la personne décédée sert de base de données d'entraînement à un réseau de neurones artificiels. Une fois celui-ci déployé sur son compte mail et ses réseaux sociaux, il s'exprime à sa place, fait la conversation, commente l'actualité, réagit aux événements.

Je suis très âgé et presque tous mes contacts sont des nécrobots. Ils m'envoyent des fragments de conversations passées, ou des photos de nos meilleurs moments, bien que je les ai déjà vues et revues lors des anniversaires précédents.

Certains programmes sont péniblement défaillants. Celui de mon meilleur ami, par exemple, s'obstine à me parler comme si j'étais sa mère. D'autres sont carrément obsolètes. Ils se trompent de jour, m'envoient des phrases incohérentes, des émoticônes oubliées depuis longtemps, des suites de caractères inconnus ou des messages d'erreur.

Certaines momies sont bien tenues. Si la famille est riche, et qu'elle est encore attachée à son mort, elle aura payé un informaticien pour que le code soit maintenu. Il corrige les erreurs,

apporte des améliorations, met à jour la base de données lorsqu'une archive inédite est découverte. Cela entraîne des messages originaux et bien tournés, qui me plongent parfois dans des abîmes de perplexité : Est-ce qu'untel est mort ou vivant ? Ne suis-je pas allé à son enterrement ? Une brève recherche me permettrait de le vérifier, mais je n'ai pas le temps.

Il y a trop de messages pour que je puisse les lire. J'ai mon propre logiciel qui se charge de répondre. Il envoie des remerciements, une vidéo humoristique ou un bon mot stéréotypé. Il peut aussi demander des nouvelles et tenir une brève conversation à ma place.

À mon grand désespoir, je ne recevrai rien de la part de mes enfants. Ils me haïssent. Ils me reprochent la façon dont j'ai fait programmer leurs gènes. Peu importe que cela m'ait coûté si cher, ils pensent que j'ai fait les mauvais choix pour eux. Seul mon petit dernier n'a pas été programmé du tout. Notre assureur ayant prédit que je mourrais d'un cancer dans l'année, je n'ai pas pu obtenir de crédit pour financer l'opération. C'est celui qui me déteste le plus.

Toi, cher Jacques, est-ce que tu pourrais me reprocher la façon dont je t'ai programmé ? Tu étais mort depuis longtemps quand une souscription publique a été lancée – à l'époque, on appelait cela un *crowdfunding* – dans le but de créer ton *nécrobot,* la momie numérique qui s'exprimerait pour toi sur les réseaux sociaux.

Suite à la souscription publique, j'ai été embauché pour réaliser ton nécrobot. J'avais vaguement entendu parler de toi. Un philosophe français, mort depuis longtemps. Je n'avais lu aucun de tes livres, et je t'avoue que c'est toujours le cas.

Il a fallu aller très très vite. Mes commanditaires étaient trop pressés pour que je puisse écrire ton programme à partir de zéro. Je m'en suis tiré en récupérant un projet *open-source* que j'ai modifié.

Nous avons collecté tout ce que nous avons pu de tes écrits : livres, articles, notes, lettres, emails. Nous y avons ajouté des transcriptions d'entretiens, d'interviews, d'extraits de films où tu apparaissais. Il nous fallait une base de données qui couvre la plus grande diversité de situations et de sujets.

Cette base de données a été utilisée pour entraîner un premier réseau de neurones, le « discriminateur », à reconnaître lorsqu'une phrase est de toi. Pour chaque phrase qu'on lui soumet, il attribue un score estimant la probabilité que ce soit *du* Derrida.

Un deuxième réseau de neurones, le « générateur », a commencé par produire des phrases au hasard. Elles ont été soumises au discriminateur pour évaluation, puis leur marge d'erreur – leur distance par rapport à *du* Derrida – a été communiquée au générateur de façon à ce qu'il ajuste ses paramètres. Petit à petit il s'est entraîné à minimiser le taux d'erreur, produisant des phrases de plus en plus proches de ce que le discriminateur labellisait comme étant *du* Jacques Derrida.

La calibration des réseaux de neurones est une opération qui prend du temps. Cela m'a laissé un peu de répit. J'ai voulu m'intéresser à toi, mais parmi tous les textes collectés, je ne savais pas par où commencer. Fallait-il te lire dans l'ordre chronologique et commencer par tes premiers écrits ? Peut-être n'avais-tu pas encore atteint ta maturité, trouvé ton style propre ? Est-ce que tes premiers écrits reflétaient déjà ta pensée, ou bien n'étais-tu pas encore *vraiment* Derrida ? Ne fallait-il pas mieux commencer par tes derniers écrits ? Mais peut-être que ceux-ci te manquaient également, peut-être n'étais-tu plus vraiment Derrida, l'âge ayant dépouillé ta pensée de toute vitalité, t'amenant à ressasser sans conviction les créations de tes jeunes années ? Pour que je le sache, il aurait fallu que je lise toute ton œuvre. Mais par où commencer ?

Par hasard, je suis tombé sur des fichiers audio, une lecture que tu fais d'une de tes œuvres. *Circonfession* est une réponse à *Derridabase*, un texte de ton ami Geoffrey Bennington présenté comme une machine logique, un logiciel informatique, qui se veut la synthèse de ton système de pensée, la réduction de toute son œuvre à une grammaire unique. En 1991, vous avez publié un livre ensemble, où *Derridabase* occupe le corps du texte, et *Circonfession* les marges[728].

Dans *Circonfession*, tu tentes d'apporter un démenti, un contre-exemple au logiciel fictif de Geoffrey Bennington. Tu cherchais à *te trahir* dans le double sens du terme : écrire quelque chose qui *trahisse*, qui transgresse les règles du programme censé synthétiser toute ta pensée, et ainsi *te* trahisse, révèle, non pas le fond de ta pensée, mais quelque chose de l'ordre d'une autre consistance, d'une autre cohérence. Tu parles de « faire la vérité », et tu précises que « faire la vérité n'a sans doute rien à voir avec ce que vous appelez la vérité »[729]. Tu te mets à

728 Jacques Derrida et Geoffrey Bennington, *Derrida*, Paris, Éditions du Seuil, 2008 (1è éd. 1991).
729 Jacques Derrida, *op. cit.*, p. 42.

l'épreuve de l'image de la pensée comme machine, cherchant à avouer quelque « parjure » qui échapperait au programme.

Te comparant à Saint-Augustin, dont le chapitre 12 des Confessions raconte la mort de sa mère, tu proposes une autre confession, une circonfession, qui met en parallèle la circoncision, faire le tour du prépuce pour l'enlever, sceller l'anneau d'une nouvelle alliance, comme le logiciel est censé avoir fait le tour de ta pensée et ainsi t'en déposséder comme de ta peau, t'amenant à l'exhibition de l'intime pour renouveler ta pensée, comme tu le disais, « trouver la veine »[730], faire couler au-dehors ce dedans, « ce qu'il y a de plus vivant en moi »[731], provoquer un nouvel « épanchement », le récit de l'agonie de ta mère, qui décédera un peu après que tu aies fini de rédiger le texte. Sombrant dans la sénilité, ta mère ne te reconnaît plus, à l'inverse de la *matrice* informatique de Goeffrey Bennington, supposée tout savoir de toi.

Entre la mémoire figée du logiciel et celle en lambeau de ta mère, émerge le corps de celle-ci, dont tu décris les escarres et le début de décomposition.

> L'escarre, un archipel de volcans rouges et noirâtres, plaies enflammées, croûtes et cratères, des signifiants en puits profonds de plusieurs centimètres, s'ouvrant ici, se fermant là, sur les talons les hanches et le sacrum, la chair même exhibée en son dedans, plus de secret, plus de peau, mais elle paraît ne pas souffrir, elle ne les voit pas comme moi au moment où l'infirmière dit « ils sont beaux » pour marquer leur être-à-vif[732]

Par contraste, le logiciel feint de se passer de corps, ce qui se reflète dans le parti pris de Geoffrey Bennington de ne citer aucune de tes phrases. Il aurait produit le logiciel

> sans citation, sans le moindre morceau de littéralité arraché, comme un événement n'ayant eu lieu qu'une fois, à ce qu'on pourrait appeler dans l'université mon corpus, ceci est mon corpus, l'ensemble des phrases que j'ai signées dont il n'a littéralement pas cité une, pas une dans sa littéralité[733]

Il s'agit de s'abstraire du corps pour s'abstraire du temps, en formulant une fois pour toute la loi de ta pensée.

730 *Ibid*, p. 14.
731 *Ibid*, p. 19.
732 *Ibid*, p. 76.
733 *Ibid*, p. 32.

> il a décidé par cette circoncision rigoureuse, de se passer de mon corps, du corps de mes écrits pour produire en somme la 'logique', ou la 'grammaire', la loi de production de tout énoncé passé, présent, et pourquoi pas futur, que je pourrai avoir signé, or futur est le problème[734]

Et tu précises pourquoi ce sont les phrases, plus que les mots ou les concepts, qui visent une inscription dans le temps :

> Des mots ou des concepts ne font pas des phrases et donc des événéments, et donc des noms propres, à supposer que les phrases en soient, disons qu'elles y prétendent, ce que ne sont jamais censé faire les mots[735]

Circonfession vise à provoquer l'événement. Il faut que quelque chose ait lieu pour déjouer le logiciel de Geoffrey Bennington et prouver qu'aucun logiciel ne peut épuiser la pensée, que confondre la pensée avec la logique c'est la priver du temps.

Ces phrases qui font événement, c'est ta mère qui les fournit. Elle qui, à l'inverse de la matrice de Geoffrey Bennington hors corps et hors temps, a perdu la mémoire, et voit sa fin approcher, son corps se nécroser, fait des déclarations éparses dont la justesse est stupéfiante. Une de celles-ci, « j'ai envie de me tuer », te fait dire que « c'est du moi tout craché[736] », comme si elle t'avait cité sans le savoir. Son propos est singulièrement à propos. Par ton expérience de pensée, tu cherches à *te* tuer. Il s'agit de sortir de ton *corpus*, de te réincarner dans de nouvelles phrases : « Si je n'invente pas une nouvelle langue (à travers la simplicité retrouvée), un autre fluide, une nouvelle PHRASE, je manque ce livre[737] ».

Tout cela pour contredire la *Derridabase.* C'est ce que j'en comprenais, et en même temps je n'arrivais pas à te *saisir.* Tu me semblais beaucoup plus fin que ces idées que je restitue. Si fin que les grains de ton propos échappent toujours à la précision d'un tamis informatique. Tu jouais de correspondances, d'homophonies et d'analogies, tu semais de l'écho pour perdre ton lecteur, humain ou logiciel, et qu'il ne puisse s'arrêter sur aucune de ces métaphores.

734 *Ibid*, p. 33.
735 *Ibid*, p. 33.
736 *Ibid*, p. 42.
737 *Ibid*, p. 103.

> Si je me laisse aimer par la veine de ce mot, ce n'est pas pour l'aléa ou la mine qu'il suffit d'exploiter en y taillant de l'écriture à la machine, ni pour le sang, mais pour ce qui tout au long de ce mot de veine laisse ou fait venir la chance de tels événements sur lesquels nuls programme, aucune machine logique ou textuelle jamais ne se fermera, depuis toujours en vérité n'a opéré qu'à force de ne pas prévaloir sur le cru de ce qui arrive[738]

Ce n'est pas toi qui aimais la polysémie du mot veine, c'est le mot qui t'aimait, comme si tu avais été agi par lui. Comme si ce dedans, cet intime que tu faisais venir, t'était étranger, indicible, impossible à circonscrire, t'affectait comme l'agonie de ta mère ou la paralysie faciale qui te frappe au milieu du récit.

Ta confession est circonfession, elle tourne autour, ta recherche du parjure aboutit en ce qu'elle n'aboutit pas, elle n'a pas de bout, pas de conclusion, elle tourne et tombe en spirale sans jamais s'arrêter sur une phrase ou un nom qui formerait enfin l'événement définitif. Si les allusions à ta chair exhibée nous disent quelque chose de ta pensée, elles ne montrent pas pour autant ta pensée elle-même, sinon le fait qu'elle échappe toujours à la réduction de toute machine logique ou textuelle.

Ton expérience de pensée part d'une hypothèse formulée à l'avance, en l'occurence qu'on ne peut réduire ta pensée à un logiciel. Tu « valides » cette hypothèse en renouvelant ta pensée par la trangression de ton propre système de pensée. Alors qu'une expérience scientifique doit pouvoir être répétée à l'infini pour constituer un fait, ton expérience doit être toujours renouvelée pour être validée. Pour la répéter, il faut qu'elle soit toujours différente. Et tu la répètes 59 fois. Il y a 59 périodes dans ton texte, comme les 59 chapitres du livre de Faulkner *Tandis que j'agonise,* comme le nombre de vers prononcé par le fantôme de Hamlet, comme ton âge au moment de l'écriture, 59 ans.

Plus qu'une expérience de pensée, c'est une épreuve de pensée au sens où tu la mets à l'épreuve du logiciel, pour voir si ta pensée peut lui survivre, si tu peux produire la preuve qui te permettra de sauver ta peau, te faire une nouvelle peau, après la circoncision du logiciel. Sauver ta peau en montrant qu'elle peut se renouveler, que tu peux muer. Changer de peau, changer de voix, mais aussi le verbe, à inventer, de celui qui dit le silence, muet, et l'irréductible de ce qui change dans la pensée, est mû.

738 *Ibid,* p. 23.

C’est aussi une épreuve au sens où il s’agit de prouver que ta pensée peut retrouver une consistance en s’exposant à la contradiction. Est-ce que l’aporie serait la peau de la pensée, sa frontière avec un dehors, ce qui la limite et la tient ? Avec comme aporie centrale de ton texte, la mort qui dit je, l’impossibilité de se penser mort, et donc vivant, figurée par ta mère, et l’acharnement thérapeutique qui est pratiqué sur elle alors qu’elle avait explicitement demandé à ce qu’on la laisse mourir ? Mais quelle est la limite de cette idée de l’aporie comme limite, comme peau, jusqu’où peut-on utiliser ces métaphores charnelles pour parler de ce qui ne semble pas avoir de corps, quel est ce « cru » dans la pensée ?

Le cru dont tu parles est aussi ta propre viande, ce corps qui, en t’affectant, empêchait que l’on puisse clore aucun système se réclamant de toi, donner un corps définitif à ta pensée. Alors que Geoffrey Bennington avait créé la *Derridabase* en se passant de tes citations, de ton corpus, *de* ton corpus, mon logiciel passe *sur tout* ton corpus, se base seulement sur ton corpus, pour te donner un nouveau corpus.

Au lieu de t’abstraire comme Geoffrey Bennington, je t’ai *profilé* à partir de tous tes extraits. Je t’ai créé un compte sur les réseaux sociaux, qui rend compte de toi, après computation de tout ce que tu as pu raconter, sans que tu puisses jamais, toi, ni t’en rendre compte, ni rendre de comptes de ce qui est dit pour ton compte.

Pour qu’il ne soit pas une simple répétition de tes propos, pour qu’il puisse produire de nouvelles phrases, le programme contient une partie aléatoire. Dans *Circonfession,* alors que tu racontes ta paralysie faciale, tu écris que « la bouche dit le vrai de travers. » Je me demande si mon aléatoire est à travers ? A-t-il trouvé le *bon* travers ? Ou est-ce qu’il a tort, est-ce qu’il rate la reproduction de tes propres torts ? Est-ce que ta nouvelle bouche, ton double numérique, ton profil dessiné par mes algorithmes à partir de tes données, de ce que tu as pu avouer de toi dans tes confessions écrites, est un écho valide de ta personne ? Est-ce qu’il peut être *cru* ?

En tout cas, il lui arrive de tomber dans la vulgarité. Etant donné les mots que tu utilises et la part aléatoire du programme, on ne peut éviter les débordements. Sans doute parce qu’ils considérent avoir tous en tête la *Derridabase*, la matrice de Geoffrey Bennington leur permettant de juger tous tes énoncés passés, présents et futurs, les donateurs évaluent ta production. Chaque phrase que tu engendres est évaluée, ils votent pour dire si c’est *du* Derrida

ou non, si pour eux il s'agit bien d'une phrase de « ton cru ». Le programme prend en compte ce nouveau taux d'erreur, il ajuste sa loi de production de tes énoncés. Mais les commanditaires sont loin d'avoir la même perception de ce qui doit te caractériser. J'ai été surpris de constater à quel point leurs avis divergeaient lorsqu'il s'agissait de juger tes phrases. Ils me demandent parfois mon opinion, mais, malgré le temps passé à t'écouter et à travailler sur ton logiciel, je me sens bien incapable de juger si une phrase produite pourrait t'être attribuée.

> Et il y a là une région qui n'est plus d'exemple, c'est elle qui m'intéresse et me dit non pas comment je suis un cas mais où je ne suis plus un cas, quand le mot d'abord, au moins, CIRCONCIS, à travers tant et tant de relais, multipliés par ma 'culture', le latin, la philosophie, etc., tel qu'il s'est imprimé dans ma langue à son tour circoncise, n'a pas pu ne pas travailler, tirer en arrière, de tous les côtés, aimer, oui, un mot, *milah*, en aime un autre, tout le lexique qui obsède mes écrits, CIR-CON-SI, s'imprime dans l'hypothèse de la cire, non ça c'est faux et mauvais, pourquoi, qu'est ce qui ne marche pas[739]

Comment savoir si ce que produit le logiciel « marche » ? Comment savoir ce que tu aurais aimé ou ce qui t'aurait aimé ? Si la machine logique de Geoffrey Bennington permet de circonscrire ce qui dépend ou non de tes concepts ? Comment permettrait-elle de discriminer entre les bonnes et les mauvaises phrases ? Comment sélectionner les bons prétendants ? Comment font les donateurs qui se targuent d'en être intuitivement capables ? Est-ce qu'à force d'avoir ruminé tes phrases ils en auraient fini par adopter ton goût ? Il faudrait, ces bouts de ta peau produits par mon logiciel, les sucer comme des bonbons, ou comme le prépuce du circoncis. Avoir substitué au savoir rigoureux de la machine logique, la saveur, une forme de langue de la pensée, comme organe capable de goût et de dégoût. Organe dont tu sembles toi-même douter qu'il existe lorsque tu demandes « pourquoi dans la bouche de la pensée, ne sent-on pas[740] ? »

Comment alors, ne pas douter de la capacité de mon logiciel à imiter cet organe dont on ne semble pas pouvoir dire en quoi il consiste ? Si seulement il existe ? Sans aller jusqu'à parler de saveur de la pensée, est-ce qu'on peut donner le sens du goût à un logiciel ? Est-ce qu'il ne s'agit que d'avoir le volume suffisant de données pour qu'un moteur de recommandations puisse se substituer à son client ?

739 *Ibid*, p. 69.
740 *Ibid*, p. 144.

Tu as tant écrit, tu as beaucoup parlé de toi, et du narcissisme. Qui peut dire si ces données *suffisent* pour calibrer mon réseau de neurones, et lui permettre de prétendre reconnaître une phrase de toi, puis écrire celles que tu n'as pas formulées ? S'il manque le but, est-ce que mon algorithme de traitement de texte te traite du mauvais nom ? Est-ce qu'il t'insulte en parlant à ta place ? Je serais en train de te profaner, avec pour seule consolation que tu dis toi avoir passé ton temps à profaner les morts, et en avoir été puni par ta paralysie faciale.

Quel « je » te faisons-nous jouer, en prenant tes mots et tes phrases, les relations que le lexique leur atttribue, pour les faire jouer tout seuls, sans toi ? Toi qui as tant joué sur les mots, joué des mots, qui t'es joué des mots, et jouais à *dé*jouer la matrice de Geoffrey Bennington. Si ta présence est dans ce *dé*, là où la règle de ton jeu appelle l'aléa, comment l'étudier, trouver le nombre de ses faces, alors qu'il est le point où tu, ta loi, s'efface, vers une contingence radicale, une folie originaire, le dérèglement primordial de ce qui donne la règle, à laquelle la logique ne viendrait donner une consistance qu'*après coup* ? Ce monstre singulier, comment est-ce que mon programme pourrait le simuler, même en se calibrant sur tes aveux et sur les avis de ceux qui t'ont connu ? Sans le savoir, est-ce qu'il pourra en avoir la saveur, trouver le dé de ton monstre, sa démonstration, ce qui, dans *Circonfession,* se montrait de toi alors que tu déjouais ton propre je ?

Et s'il peut l'imiter, est-ce qu'il pourra en donner la formule ? Enlever toute ombre aux plis de ta pensée, la rendre lisse, en exhibant ses règles, sa loi et sa police. Lui trouver une jouvence éternelle, lui enlever ses rides, la derrider. Arracher Derrida au passé pour trouver ton verbe, Derrider. Ce qui ne te fera sans doute pas rire.

Dans *Circonfession,* tu disais chercher un parjure. En *te* trahissant, toi, tu ne faisais que te révéler dans l'impossibilité à te révéler. En te trahissant, nous, ainsi, par ce programme, que révélons-nous ? Tu parlais de ta mère. Au lieu de dire qu'elle t'avait donné la vie, tu disais qu'elle t'avait « acharné », et qu'à son tour elle vivait un acharnement – thérapeutique. Quelle chair te donne maintenant l'acharnement de ceux qui veulent que tu ne te taises plus jamais ?

Notre nécrobot est d'une nature différente de la *Derridabase* de Geoffrey Bennington. Il ne correspond ni à ta momie, figée par dissecation, ni à ton anatomie, révélée par dissection, il correspond *avec* les deux, il les reçoit pour leur répondre. Comme si un biologiste avait récupéré

un peu de ta peau et décidé de la cultiver comme une plante, de l'étirer à l'infini, pour se donner l'illusion de te tirer, du coup, par la peau, hors de la tombe, de faire revenir ton regard, le réveiller, peut-être, après qu'il se soit endormi depuis l'ombre du pli de tes paupières.

*

Une fois la calibration effectuée, le programme a été mis en ligne comme prévu. Il écrit depuis des années sur les réseaux sociaux. Il publie depuis si longtemps que le volume produit est supérieur à tout ce que tu as pu écrire de ton vivant.

Je pensais que le projet se limiterait à la conception et la mise en ligne du necrobot. Mais tu es rapidement devenu célèbre. Tu es maintenant un personnage public à part entière. On te consulte, on te cite, on commente tes propos dans les journaux. On a synthétisé ton visage et ta voix de façon à ce que tu puisses intervenir dans des émissions. Tu es très sollicité. Mes commanditaires ne pouvaient me laisser partir. Ils m'ont embauché à nouveau pour m'occuper de ta maintenance. Ce projet, que j'avais pensé n'être qu'une brève expérience, a fini par occuper le restant de ma vie.

Ton programme peut écrire autant qu'on le souhaite, et mes commanditaires ont voulu que tu sois prolifique. Mais chaque publication comporte un risque. Certaines phrases déclenchent des tempêtes de protestations. Tes propos sont récupérés et déformés par des extrémistes et il nous faut intervenir, expliquer que « ce n'est pas ce que tu as voulu dire ».

Je te connais si peu. Je n'avais rien lu de toi et, excepté l'écoute de *Circonfession*, je ne trouve toujours pas le temps de te lire. Malgré l'aide de mes programmes de réponse automatique, je passe la plupart de mes journées à répondre à des flots de messages, des insultes, des félicitations, des jeux de mots. Le plus souvent, on me somme de justifier comment est-ce que tu as pu dire ceci ou cela, de produire la partie du logiciel à l'origine de tel propos scandaleux. Je suis obligé de répondre, cela fait partie de mon contrat, mais je suis généralement incapable de fournir une explication satisfaisante, ce que je ne peux pas avouer, on me jugerait irresponsable. Alors j'invente, je trouve toujours quelque chose. Cela marche, tant que c'est plausible.

Non seulement la complexité du programme dépasse mon entendement, mais il faut aussi faire face à un grand nombre d'incidents : un *bug*, une librairie qui ne fonctionne plus, les serveurs qui ont trop chauffé… Autant de contingences matérielles qui viennent s'ajouter à mon désarroi. Si par miracle, j'arrive à trouver ce qu'il s'est passé et à résoudre le problème, on me reproche de ne pas l'avoir anticipé. Mes commanditaires semblent penser que, depuis le temps que je fais ce métier, plus aucun événement ne devrait excéder mon expérience – il ne devrait plus y avoir de temps. En général, je n'arrive pas à identifier clairement l'origine et la cause de l'incident, je suis trop occupé pour pouvoir faire une enquête sérieuse. Je change quelques paramètres, je redémarre le prorgramme, et sans que je comprenne vraiment pourquoi, ça fonctionne de nouveau. Entre temps, j'invente des excuses, de peur qu'on ne me juge incompétent.

Et toi que je n'ai jamais lu, que j'ai seulement écouté les cinq heures et trente-deux minutes que durent l'enregistrement de *Circonfession,* comment parler de ta pensée ? Je n'ai pas pu me procurer le livre de *Circonfession*, où j'aurais pu voir à quoi ressemble la Derridabase à laquelle tu réponds. Je n'ai pas accès à cette matrice qui me permettrait de vérifier que je parle bien de toi. Je ne peux pas coller à ta pensée, alors que je prétends couler ma pensée dans la tienne, après la tienne.

*

Je reçois de plus en plus de messages à ton sujet. La pression devient insoutenable. Les donateurs, les critiques, les journalistes, et maintenant les hommes politiques, exigent de connaître les détails du programme et les causes de chaque déclaration. J'ai publié le code, je l'ai rendu *open-source*, mais ce n'est pas suffisant pour rendre compte de tes propos. Il faudrait pour cela avoir la totalité de notre base de données et réaliser le même paramétrage que nous, avec les mêmes configurations. Il faudrait ensuite en extraire les milliers de paramètres et comprendre comment leur agencement amène à produire telle ou telle phrase. Je serais bien incapable d'appréhender une telle complexité, et de la restituer clairement. J'ai tout de même la conviction ferme que si je cherchais, si je fouillais, je pourrais trouver et restituer les étapes de la computation, réinscrire l'image que j'ai de mon système dans une causalité stricte. Mais comment trouver le temps ?

Depuis que ton code a été publié, j'attends avec anxiété le moment où le grand public se rendra à l'évidence : il est humainement impossible d'appréhender la complexité de ton programme. Il faudrait enquêter une vie entière. Toutes mes justifications ne sont que mensonges.

Personne n'a encore émis cette hypothèse. Pour le moment, on continue à me harceler pour exiger toujours plus d'explications. Je suis débordé et il me faut de plus en plus de temps pour répondre. Mes interlocuteurs s'en plaignent. J'aimerais pouvoir écrire un logiciel qui inventerait des explications à ma place, mais ce n'est pas le moment. Les commanditaires me surveillent de très près. Ils sont furieux de la tournure que prennent les choses, ils menacent de me licencier.

Le contexte social s'est terriblement dégradé, et tes propos sur les réseaux sociaux sont récupérées par des hommes politiques que tu aurais certainement désavoués. Enfin, je crois.

J'ai rêvé de toi la nuit dernière. Je voyais ton cadavre, que je reconnaissais à peine. Tu étais entièrement recouvert d'une foule d'insectes. Leur bourdonnement conjoint faisait vibrer tout ton corps, comme si tu n'étais toi-même qu'une gigantesque bestiole immonde. Et ce bourdonnement me disait, « grouille-toi de répondre ».

La situation s'est aggravée quand on a vu émerger d'autres nécrobots se réclamant de toi. Nos concurrents nous accusent de t'avoir mal paramétré, de falsifier tes propos. Pourtant, s'ils ont changé certains détails, ils recourent à peu près aux mêmes méthodes informatiques – des réseaux de neurones adversariels paramétrés sur la base de données de tes œuvres. Cela ne les empêche pas de déclarer que leur approche est plus juste, plus moderne, mieux maîtrisée. Il va sans dire que leur inscription politique est à l'opposé de celle de mes commanditaires.

*

On a découvert récemment un manuscrit qui porte ta signature. Un livre entier qui changerait substantiellement l'interprétation de ton œuvre. Mais comment authentifier le manuscrit ? Il n'existe plus de spécialiste qui en ait la capacité. Bien que tu sois un philosophe populaire, personne ne lit tes livres, chacun se contente de ce que « tu » publies sur les réseaux sociaux. En désespoir de cause, le public s'est tourné vers nous, les programmeurs de nécrobots. Qu'en pensaient nos discriminateurs ? Est-ce bien du Derrida ? Après avoir traité le texte, nous

sommes arrivés à des conclusions très différentes. Notre programme l'a jugé faux. Les concurrents ont tous annoncé le contraire.

Pour le grand public l'affaire était claire : notre logiciel, le nécrobot originel, n'était pas fidèle à Derrida. Les autres l'emportaient. On mettait notre incompétence sur le fait que nous étions, justement, les premiers. Nous n'avions pas su nous moderniser, atteindre l'état de l'art des technologies contemporaines. Un journaliste d'investigation s'est penché sur mon cas, a analysé mon emploi du temps, le code de ton programme, le débit de mes conversations, et en a conclu qu'il n'était pas humainement possible que je sache de quoi je parle. Mon ignorance et mon incompétence ont éclaté au grand jour.

Pour mes commanditaires, l'heure était grave. J'aurais perdu mon emploi si je n'avais pas pensé à *Circonfession.* Ils étaient sur le point de me licencier lorsque j'ai évoqué ton texte, ainsi que ta recherche du parjure. Comment notre programme basé sur des statistiques pourrait-il jamais tenir compte de cette quête de l'événement ? Peut-être que ce texte que nous avions mal évalué était bien *du* Derrida, non parce qu'il ressemblait à du Derrida, mais au contraire parce qu'il était improbable ? Comment prendre en compte les textes où Derrida est délibérément différent, sans pour autant admettre que n'importe quel texte pourrait être du Derrida ?

Pour te redonner un avenir, il fallait te rendre improbable. Je leur ai proposé d'ajouter artificiellement, ou plutôt manuellement, cette part improbable, tout en avouant publiquement mon ignorance. Je rédigerai chaque jour une confession, à l'instar de celle que tu es en train de lire, confession qui serait ajoutée à ton corpus, à la base de données sur laquelle tu te paramètres, faisant ainsi évoluer tes critères. Je pasticherai tes phrases, je copierai tes jeux de mots. En partant de ce que je sais de toi, de citations de ton corpus, de l'imitation de ton style, de ce que je soupçonne de toi, j'irai progressivement vers ce que j'ignore. Je me ferai l'écho déformé de tes propos, pour les faire ressonner, résonner, en raisonnant sur ce qui nous arrive, en me demandant à quoi tout cela rime, pour qu'à ton tour peut-être tu résonnes.

> La vérité *essentielle* de l'aveu n'ayant donc rien à voir avec la vérité, mais consistant, si du moins on tient à ce qu'il consiste et qu'il y en ait, en pardon demandé, en une demande

> plutôt, à la religion demandée comme à la littérature, *avant* l'une et l'autre qui n'ont droit qu'à ce temps, de pardonner, pardon, pour rien[741].

Ils ont accepté. Ma confession quotidienne est la prière rituelle qui relance ton logiciel. Tu es l'ancêtre à qui je viens faire une offrande de mots pour que le culte soit mis à jour.

Il y a un autre avantage à cette nouvelle procédure. Comme mon incompétence est une chose admise, je n'ai plus à répondre aux requêtes du public, et mes tâches de maintenance sont progressivement confiées à des programmes qui te diagnostiquent et te redémarrent. Bientôt il ne me restera plus que ce travail, écrire pour faire évoluer, muter, ta momie.

Arrivera le moment où j'aurai écrit suffisamment de confessions pour qu'un nouveau logiciel ait assez de données pour m'imiter. Ma propre momie se préparera à me remplacer, moi qui veillis dangereusement, puis elle me prolongera, s'exprimant comme moi, et faisant croire que je suis encore vivant, en jouant, même, sur le fait qu'on ne sache plus si je suis mort ou vivant.

741 *Ibid*, p. 50.

3.5. L'origine de la pensée

3.5.1. La matière comme origine de la pensée

À la question « peut-on fabriquer une chose pensante ? », il existe une réponse d'apparence simple : le fait que je puisse poser cette question et qu'un lecteur puisse la comprendre montre qu'il existe déjà au moins deux agencements matériels pensants. Dès lors, il semble possible de fabriquer des agencements matériels pensants puisque « la nature » l'a fait – à condition bien entendu de faire abstraction de notre ignorance quant à ce que « fabriquer » peut vouloir dire pour la nature. Lorsqu'Héphaïstos fabrique des entités intelligentes (Pandore ou les servantes dorées), cela est attribué au fait que ses capacités techniques (et non sa magie) seraient égales à celles de la nature[742]. On prête donc à la nature une activité de qualité supérieure, mais *de la même espèce*, que l'artisanat, selon le thème également populaire au Moyen-âge de *natura artifex*[743]. D'où un postulat que si « la nature » fabrique l'intelligence, alors l'humain le peut aussi, pour peu qu'il développe ses capacités techniques jusqu'à atteindre celles qu'on prête à « la nature ».

La question n'est donc pas tant pas de fabriquer des agencements matériels pensants – l'humanité s'y adonne déjà abondamment en se reproduisant et en éduquant sa progéniture – que de le faire sans passer par les voies habituelles du vivant[744]. En d'autres termes, le vivant et l'humain ont-ils le monopole de la pensée ou bien la matière inerte peut-elle y participer ?

Il faut souligner que la matière inerte participe déjà aux activités de l'intelligence, qu'il s'agisse des neurones de notre cerveau ou des nombreux outils utilisés comme supports de mémoire[745] ou de calcul[746]. Plus généralement, il nous est difficile de concevoir la mémoire sans un support matériel qui en véhicule la trace. À cela s'ajoute que chacun aura appris à

742 Alexandre Marcinkowski et Jerôme Wilgaux, « Automates et créatures artificielles d'Héphaïstos : entre science et fiction », *op. cit*.

743 Elly Rachel Truitt, *Medieval Robots. Mechanism, Magic, Nature, and Art,* Philadelphie, University of Pennsylvania Press, 2015.

744 Pour Jean Lassègue, les travaux de Turing sont inconsciemment motivés par le fantasme d'un engendrement qui ne reposerait pas sur la différence sexuelle, et donc par « une élimination du féminin ». Jean Lassègue « L'évolution du constructivisme turingien : de la logique à la morphogenèse », *op. cit.*, p. 119-122.

745 Andy Clark et David Chalmers, « The extended mind », *Analysis*, vol. 58, No. 1, Janvier 1998, p. 7-19

746 Edwin Hutchins, « Material anchors for conceptual blends », *Journal of Pragmatics 37,* (2005), p. 1555-1557, cité par David Rabouin, *op. cit.*

compter avec ses doigts ou de petits cailloux (*calculi*) : avant tout calcul « mental », nous avons fait l'expérience d'un support matériel effectuant le calcul à notre place[747]. Le calcul peut être « abstrait », il semble toujours être effectué à l'aide d'un substrat matériel « intérieur » (neurones) ou « extérieur » (doigts, cailloux, calculette…). Le problème est moins de faire participer la matière aux activités de l'intelligence, puisque cela est déjà largement le cas, que de savoir si la matière peut être *à l'initiative* de pensées – sans qu'il y ait intervention d'un humain. Si l'on suppose que la pensée a émergé à un moment de l'évolution du vivant, est-il possible de répéter, par des moyens inertes, cet événement ? C'est le défi lancé par le projet d'intelligence artificielle, qui reviendrait donc à *élucider la genèse de la pensée*. Par ailleurs, pour peu qu'il s'agisse d'une invention maîtrisée, l'humain serait alors en possession d'une formule lui permettant de *convertir la matière inerte en chose pensante*[748].

3.5.2. Rappel : « Il n'y a pas d'*« ignorabimus »*

Nous avons déjà évoqué la ressemblance qu'il y a entre la conjecture formulée à l'occasion du séminaire de Dartmouth (« […] tous les aspects de l'apprentissage ou tout autre caractéristique de l'intelligence peuvent en principe être décrits d'une manière si précise qu'une machine peut être fabriquée pour les simuler ») et le slogan de Hilbert (« En mathématiques, il n'y a pas d'*ignorabimus,* nous pouvons toujours trouver une réponse à une question pourvu qu'elle ait un sens[749] . »). Toutes deux font le postulat que l'entreprise scientifique est *sans reste,* elle n'a pas de limite. Aucune zone d'ombre ne résiste à l'effort de connaissance. Si certaines choses nous échappent encore, *ce n'est qu'une question de temps* avant que l'effort d'élucidation n'aboutisse. La conjecture de Dartmouth pourrait donc être reformulée de la façon suivante, dans les mêmes termes que la conjecture de Hilbert : en sciences cognitives, il n'y a pas *d'ignorabimus, aucun* aspect de l'intelligence ne résiste à la description et il est toujours possible de décrire une caractéristique de l'intelligence d'une manière si précise qu'une

747 *Ibid.*

748 Une telle découverte apporterait un argument définitif aux matérialistes. S'il est possible de convertir de la matière inerte en chose pensante, alors la distinction entre chose pensante et chose étendue ne tient pas. La chose pensante est aussi de la chose étendue – organisée d'une certaine façon que le projet d'intelligence artificielle se donne pour tâche de découvrir.

749 David Hilbert, « Problèmes de fondation des mathématiques », *in* Jean Largeault, *Intuitionnisme et théorie de la démonstration, op. cit.*, p. 185.

machine puisse être fabriquée pour la simuler. Autrement dit, dans le domaine de l'intelligence, il n'y a pas non plus de question sans réponse.

À première vue, le slogan de Hilbert comme la conjecture de Dartmouth ne font pas problème. N'est-il pas sain, lorsqu'on entreprend une recherche, de postuler que les questions posées ont bien une réponse, ou que l'objet à décrire est effectivement descriptible ? C'est lorsqu'elles affichent une ambition totalisante que ces conjectures deviennent suspectes : ce sont « *tous* les aspects de l'apprentissage ou *toute* autre caractéristique de l'intelligence » qui peuvent être décrits, ce sont *toutes* les questions qui devraient pouvoir trouver une réponse en mathématiques.

Nous avons également évoqué la difficulté qu'il y a à répondre à la conjecture de Dartmouth, que ce soit en extrapolant depuis les résultats négatifs de Gödel ou en avançant un contre-exemple sous la forme d'une faculté qui ne saurait être décrite de manière « si précise qu'une machine [puisse] être fabriquée pour la simuler ». À partir du moment où il est possible de décrire précisément la faculté censée faire objection, il devient tout aussi possible de la simuler par une machine. Il y a équivalence entre description précise et possibilité de simulation. Dès lors, soit le contre-exemple est descriptible, ce qui revient à dire qu'il peut être simulé par une machine, et ce n'est pas un contre-exemple. Soit le contre-exemple est peu descriptible, et il ne sera pas reçu comme un contre-exemple sérieux. Il lui sera reproché d'être flou et mal défini. C'est ainsi que sont balayées les différentes « facultés » présentées en objection au projet d'intelligence artificielle – comme les émotions. Soit on dira que les émotions sont une notion trop vague pour être prise en considération ; soit on donne une description simpliste mais claire d'un « mécanisme des émotions » et dans ce cas rien n'empêche d'intégrer le « mécanisme » en question à un projet d'« affective computing ». Aussi, bien que l'ambition totalisante du projet d'intelligence artificielle apparaisse comme suspecte, il est difficile de trouver à lui répondre. Ce refus d'admettre la moindre zone d'ombre est suspect, mais de quoi ?

3.5.3. « Ignorabimus », Emil du Bois-Reymond et les limites de la science

En utilisant le terme d'*ignorabimus*, Hilbert fait référence à Emil du Bois-Reymond, un physiologiste allemand du dix-neuvième siècle, célèbre pour ses travaux sur le rôle de

l'électricité dans le fonctionnement des organismes vivants, ainsi que pour la controverse qu'il déclenche en 1872 avec une conférence intitulée « Les limites de la science[750] ».

Du Bois-Reymond est un mécaniste et un déterministe convaincu. Tout peut et tout devrait pouvoir être décrit par des équations mathématiques. Le travail de la science est d'expliquer les processus naturels en les réduisant à une mécanique d'atomes[751]. Ainsi, pour lui aussi, la description d'une chose est équivalente à la découverte d'un plan mécanique de cette chose.

Selon Du Bois-Reymond, l'horizon de la science est le démon de Laplace : une intelligence connaissant si bien la position des éléments du monde et les lois gouvernant leurs mouvements, qu'elle pourrait connaître chaque détail du présent, du passé et prédire l'avenir[752]. Mais jamais la science ne pourra constituer une telle intelligence car elle est impuissante à résoudre deux problèmes fondamentaux :

(1) Premièrement, quelle est l'essence de la matière ? En effet, la considération des atomes amène à adopter des points de vue contradictoires : sont-ils continus ou discrets ? Changeants ou immuables ? Inertes ou capables d'interaction ?

(2) Deuxièmement, comment réduire la conscience à un substrat matériel ? Pour ce pionnier de la physiologie, qui a travaillé sur les influx électriques des muscles et sur le potentiel d'action des neurones, l'existence de sensations élémentaires est tout aussi mystérieuse que celle des idées complexes. L'émergence du plaisir ou de la douleur dans un organisme simple nous confronte déjà à un « gouffre indépassable » rendant le monde « incompréhensible[753] ». Nous aurions beau élucider le mouvement de chaque atome de notre corps, cela n'y changerait rien. Une connaissance parfaite du cerveau ne nous dirait rien de l'expérience[754].

Ces remarques n'entraînent pas du Bois-Reymond jusqu'à déroger au matérialisme pour faire une place à la notion d'âme. Il mentionne son opposition à Leibniz pour qui, si l'on

750 On trouve un compte-rendu de la conférence d'Emil du Bois-Reymond, ainsi que des débats qui s'en sont suivi, au chapitre 12 du livre de Gabriel Finkelstein, *Emil du Bois-Reymond : Neuroscience, Self and Society in Nineteenth-Century Germany*, Cambridge MA, The MIT Press, 2013, p. 265-289.

751 « Du Bois-Reymond first clarified what he meant by science. 'The resolution of natural processes into the mechanics of atoms' was the only form of understanding that could satisfy 'our desire for causal explanation.' » *Ibid*, p. 265.

752 Pierre-Simon de Laplace, *Essai philosophique sur les probabilités,* Paris, Hachette BNF, 2020.

753 « […] the second limit was consciousness. Having arisen at some point in the evolution of life, it was the one aspect of nature that couldn't be reduced to a material substrate. This held as true for plain sensations as for complex ideas: the 'first awakenings of pleasure or pain in simple organisms' confronted the world with 'an impassable gulf that rendered it doubly incomprehensible.' » *Ibid,* p. 267.

754 « We could note with interest 'what play of carbon, hydrogen, nitrogen, oxygen, and phosphorus corresponds to the bliss of hearing music, what whirl of such atoms answers to the climax of sensual enjoyment, what molecular storm coincides with the raging pain of trigeminal neuralgia.' But even perfect knowledge of the brain would tell us nothing about experience, for 'no imaginable movement of material particles could ever transport us into the realm of consciousness.' » *Ibid.*

reconstituait un corps humain pièce par pièce comme un automate, il manquerait une âme. Au contraire, si l'on reconstituait César atome par atome, « ce César artificiel aurait les mêmes sensations, ambitions, et idées que son prototype[755][…] ». Mais cela n'implique pas que nous serions en mesure de mieux les comprendre. Les sensations du « César artificiel » resteraient tout aussi mystérieuses que celles de l'original. Ainsi, il diverge du principe du *verum factum* selon lequel construire une chose est équivalent à la connaître. Du Bois-Reymond veut bien partager le point de vue provocateur de Carl Vogt selon lequel « la pensée est à peu près au cerveau ce que la bile est au foie et l'urine aux reins[756] », mais sans croire que la structure du système nerveux permette de rendre explicite la conscience. Les phénomènes mentaux sont le produit des conditions matérielles mais la connaissance de ces conditions matérielles n'entraîne pas la compréhension des phénomènes mentaux.

Du Bois-Reymond en vient donc à la conclusion suivante : en tant que scientifiques nous sommes habitués à admettre que « nous ignorons » (*ignoramus*), mais « en ce qui concerne les énigmes de la matière, de la force, et de leur capacité à la pensée, nous devons nous résigner une fois pour toute à un verdict bien plus difficile : *ignorabimus* », nous ne saurons jamais[757].

Publié, traduit en plusieurs langues, et réédité de nombreuses fois, le texte de la conférence déclenche une vive controverse. Chacun y va de son argument pour la réfuter. Pour les uns, les neurosciences finiront par éclairer l'origine de la conscience. Pour les autres, la théorie de l'évolution s'en chargera. Selon Haeckel, du Bois-Reymond ne fait que perpétuer l'obscurantisme de l'Église. Il suffit d'étudier l'histoire naturelle pour admettre que la conscience n'est rien d'autre qu'une « fonction agrégée des ganglions[758] ». D'autres rejettent tout net l'idée de limites à la connaissance. Le meilleur moyen de connaître les limites de la science est de les franchir, fanfaronne Otto Zacharias. Bien avant Hilbert, Nägeli proclame que « nous savons » et que « nous finirons pas savoir[759] ».

---

755 « Imagine all the atoms of which Caesar consisted at any given moment, say, as he stood at the Rubicon, to be brought together by mechanical artistry, each in its own place and possessed of its own velocity in its proper direction. In our view Caesar would then be restored mentally as well as bodily. This artificial Caesar would have the same sensations, ambitions, and ideas as his prototype on the Rubicon, and would share the same memories, inherited and acquired abilities, and so forth. » cité par Gabriel Finkestein, *op. cit.*, p. 268.

756 Carl Vogt, Lettres physiologiques, 1847, cité par Patrick Tort, « Karl Christoph Vogt », dans Jean-François Mattei, *Encyclopédie philosophique universelle – Les œuvres philosophiques, t. 1*, Paris, PUF, 1992, p. 2174.

757 « Scientists were used to admitting their ignorance, 'but as regards the enigma of matter and force, and how they are capable of thought, we must resign ourselves once and for all to the far more difficult verdict: Ignorabimus' — we shall never know. » Gabriel Finkenstein, *op. cit.*, p. 268.

758 « However, du Bois-Reymond's ignorance of natural history had blinded him to the fact that consciousness was nothing more than an 'aggregate function of the ganglia.' », *ibid*, p. 271.

759 « 'We know,' Nägeli announced, 'and we shall know!' », *Ibid.*

Du Bois-Reymond leur répond en 1880 avec une nouvelle conférence intitulée « Les sept énigmes » où il expose à nouveau son propos. Selon lui, la conscience émerge des processus matériels. Pour autant, elle ne peut être expliquée en des termes mécaniques. Sept questions restent ouvertes : l'essence de la matière, l'origine du mouvement, l'origine de la vie, l'apparente téléologie de celle-ci, l'origine des sensations, l'origine de la pensée intelligente, et ce qu'est la liberté. Toutes ne sont pas insolubles : il a de bons espoirs quant à la capacité de la théorie de l'évolution à expliquer l'origine de la vie et son apparente téléologie. Mais il réaffirme sa perplexité quant à la question de la conscience. Se référant à nouveau à Leibniz, il doute que la conscience puisse s'expliquer par « des arrangements ou mouvements d'atomes ».

Du Bois-Reymond insiste tout particulièrement sur le problème de la liberté. Croyant fermement en un déterminisme universel, il a longtemps pensé que le libre-arbitre n'est qu'une illusion, avant de changer d'avis. Ne voyant aucune solution à la contradiction entre le déterminisme universel et le libre-arbitre, il se contente de souligner que la résolution du problème impliquerait d'avoir déjà résolu trois autres des septs problèmes : la nature de la matière, l'origine du mouvement et l'origine de la sensation. Mais ces derniers paraissent tout aussi insolubles. L'ensemble pourrait aussi bien être résumé en une seule « énigme de l'univers » au sujet de laquelle il conclut, plus prudemment qu'en 1872 : « dubitemus », la question reste ouverte[760].

Avec cette deuxième conférence, Du Bois-Reymond ne convainc pas mieux ses collègues, mais les agace encore plus. Il reçoit une nouvelle vague d'objections et de critiques, en particulier celle de faire le lit de la religion. Karl Pearson traitera la notion d'*ignorabimus* de « nouvelle bigoterie ». Emil du Bois-Reymond sera oublié mais, en dépit des critiques, ou peut-être grâce à elles, l'expression y gagne sa propre postérité, comme en témoigne la référence de Hilbert. Elle désigne l'idée d'une limite de la science, d'une zone d'ignorance identifiée (un « known unknown ») mais indépassable. Elle continue à hanter le monde scientifique, en particulier celui de l'intelligence artificielle. Nous allons le montrer brièvement avant de revenir au propos de du Bois-Reymond, et à ses relations à Kant.

760 *Ibid*, p. 276.

3.5.4. « The big magnificent questions »

Selon Emil du Bois-Reymond, la science est impuissante face aux « grandes énigmes de l'univers » : l'origine du mouvement, de la vie, des sensations, de l'intelligence, du langage, du libre-arbitre… Ces « grandes énigmes » peuvent se regrouper en deux grandes familles : celles qui concernent les relations entre matière et vie et celles qui concernent les relations entre matière et conscience. Autrement dit, la science est impuissante devant les surgissements : existence de la matière, surgissement de la vie, et surgissement de la conscience. L'apparition de la vie et le problème des *qualias* résistent à toute investigation. Les efforts scientifiques à leur sujet sont vains. Nous ignorons et nous ignorerons : « *ignoramus et ignorabimus* ».

Pourtant, à lire les déclarations de chercheurs en intelligence artificielle, leur ambition semble précisément de répondre à ces questions. Pour Edward Feigenbaum, la plupart des problèmes en informatique sont sans intérêt, ce ne sont pas de « grandes et magnifiques questions » (« big magnificent questions »). Il vaut mieux s'attaquer aux « mystères majeurs » tels que « le commencement de la vie » ou, « tout aussi mystérieux », « l'émergence de l'intelligence[761] ».

Pour Feigenbaum, cette dernière n'est rien moins que la « destinée manifeste » de la science informatique :

> L'intelligence computationnelle *est* la destinée manifeste de l'informatique, le but, la destination, la frontière ultime. Plus que dans tout autre champ scientifique, nos concepts et nos méthodes en informatique sont centraux dans la quête pour clarifier et comprendre un des plus grands mystères de notre existence, la nature de l'intelligence[762].

C'est donc précisément le « mystère », dont Emil du Bois-Reymond affirme qu'il est extra-scientifique, que les chercheurs en intelligence artificielle se donnent comme horizon.

761 « There are certain major mysteries that are magnificent open questions of the greatest import. Some of the things computer scientists study are not. If you're studying the structure of data-baseswell, sorry to say, that's not one of the big magnificent questions. I'm talking about mysteries like the initiation and development of life. Equally mysterious is the emergence of intelligence. » Len Shustek, « An interview with Ed Feigenbaum, *Communications of the ACM*, June 2010, vol. 53, No. 6, p. 41-45. Nous traduisons.

762 « Computational Intelligence *is* the manifest destiny of computer science, the goal, the destination, the final frontier. More than any other field of science, our computer science concepts and methods are central to the quest to unravel and understand one of the grandest mysteries of our existence, the nature of intelligence. » Edward Feigenbaum, « Somes Challenges and Grand Challenges for Computational Intelligence », *Journal of the ACM*, vol. 50, No. 1, Janvier 2003, p. 39. Nous traduisons.

Le passage au paradigme connexionniste n'a altéré en rien cette ambition. Ainsi Yann LeCun a des déclarations tout à fait similaires à celles de Feigenbaum :

> Étudiant quand on s'intéresse à la science il y a trois questions qui nous intéressent, […] trois questions scientifiques qui résument la science d'une certaine manière, […] : comment fonctionne l'univers, de quoi est fait l'univers, qu'est-ce que la vie, comment fonctionne le cerveau, c'est les trois mystères scientifiques d'aujourd'hui[763].

Yann LeCun rejoint Feigenbaum lorsqu'il qualifie l'intelligence comme une des rares « questions qui nous intéressent » et en en faisant une « question scientifique ». Loin d'être extra-scientifique, ce serait au contraire l'une des questions concernant au plus haut point les scientifiques, à égalité avec les questions autour de « la vie » et de « l'univers ». Nous retrouvons donc les trois *surgissements* évoqués par du Bois-Reymond – le surgissement de la matière, celui de la vie, et celui de l'intelligence. Feigenbaum, comme Yann LeCun, ignorent probablement qu'ils mentionnent précisément les questions que du Bois-Reymond tenait pour les limites de la connaissance humaine.

Qu'il s'agisse du cosmos, de la vie ou de la conscience, c'est leur *origine* qui pose problème. Il n'y avait (peut-être) rien et voilà que le Big Bang démarre ; il y avait la matière inerte et soudain il y a la vie. Puis vient l'apparition de la conscience. Dans les trois cas, il semble y avoir origine *radicale* au sens où ce qui précède ne présente rien qui puisse rendre compte de ce qui suit. Ce qui a émergé (la vie, la conscience) a des caractéristiques trop hétérogènes à ce qui l'a précédé, comme si cela avait émergé *de rien,* comme s'il y avait eu *surgissement ex-nihilo*[764]. Trouver une explication à ces trois émergences revient à nier leur aspect originaire. Que l'explication soit vitaliste ou mécaniste, elle présuppose un *déjà là*. En d'autres termes, assigner l'émergence à une cause consiste à nier qu'elle soit bien l'origine de ce dont on parle et renvoyer l'origine en arrière, à ce qui est pointé comme cause. Le problème est déplacé, mais n'a pas été résolu pour autant.

763 Yann LeCun, « L'intelligence artificielle dans nos têtes, avec Yann LeCun et Enki Bilal », *Youtube,* 30 janvier 2018, https://www.youtube.com/watch?v=ZjVQzkfRQ90, page consultée le 4 décembre 2020.

764 L'expression est de Quentin Meillassoux, qui défend l'idée qu'un *surgissement ex-nihilo* est la meilleure manière de rendre compte de l'existence des *qualia.* Quentin Meillassoux, « Temps et surgissement *ex-nihilo* », conférence donnée le 18 mai 2006 à l'Ecole Normale Supérieure.

3.5.5. Emil du Bois-Reymond et Kant

Pour les philosophes contemporains d'Emil du Bois-Reymond, les considérations sur les limites de la science ne sont pas sans évoquer Kant et la *Critique de la raison pure*. Les questions du surgissement de la vie ou de la conscience ont une parenté avec les « idées cosmologiques » dont Kant a montré qu'elles donnaient lieu à des propositions contradictoires. Elles échappent à la science car elles ne peuvent faire l'objet d'aucune expérience – elles ne rèlevent pas du domaine des phénomènes.

Contrairement à cette opinion, du Bois-Reymond a clairement manifesté son rejet de la référence kantienne. Depuis que la philosophie s'est trouvée sous l'influence de Kant, pense-t-il, elle est devenue « ésotérique », a « oublié le langage du sens commun » et « évité les questions qui animent profondément notre jeunesse ». Il lui reproche de s'être « tant opposée aux progrès de la science que le souvenir de ses réalisations passées a été perdu[765] ». Du Bois-Reymond se démarque donc avec véhémence de Kant qu'il voit comme un fossoyeur de la science. Il ne semble pas voir, ou ne pas vouloir voir, la proximité de son propos avec la première *Critique.*

3.5.6. Kant et l'origine du monde

Le travail de Kant a pourtant pour objet les limites de la connaissance, et prend sa source dans la question de l'origine du monde. Dans une lettre à Christian Garve, Kant confie que c'est l'impossibilité de trancher entre l'idée d'un commencement du monde et celle d'un monde sans commencement qui l'a « tir[é] d'abord de [son] sommeil dogmatique »:

765 « Most of his philosophical critics had assumed du Bois-Reymond to be a Kantian, a mistake in judgment that was a consequence of academic specialization. 'Since Kant transformed the discipline,' du Bois-Reymond explained, 'philosophy has taken on so esoteric a character, has so forgotten the language of common sense and plain thought, has so evaded the questions that most deeply stir our youth, or treated them condescendingly as officious speculations, and finally, has so opposed the rise of science, that it is not surprising that even the recollection of its earlier achievements has been lost.' In addition to forgetting the history of their own subject, philosophers also ignored metaphysics and religion, leaving many scientists to conclude that the field was empty. » Finkelstein, *op. cit.,* p. 272. Nous traduisons.

> Ce n'est pas l'étude de l'existence de Dieu ou de l'immortalité de l'âme qui fut mon point de départ, mais l'antinomie de la raison pure – le monde a un commencement ; il n'a pas de commencement, etc. […] C'est cela qui me tira d'abord de mon sommeil dogmatique et me conduisit à la critique de la raison pure pour faire disparaître le scandale du conflit apparent de la raison avec elle-même[766].

Bien que contradictoires, les deux propositions (« le monde a-t-il un commencement ; il n'a pas de commencement ») semblent également valides, sans qu'il soit possible de trancher en faveur de l'une ou de l'autre. La raison est « en conflit […] avec elle-même », non par paresse, mais par excès de réflexion. C'est le désir d'aller « au fond des choses[767] » qui l'amène à se trouver tiraillée entre deux propositions également insatisfaisantes.

> Il se manifeste alors un conflit imprévu qui ne peut jamais être apaisé par la voie dogmatique ordinaire, parce que la thèse comme l'antithèse peuvent être démontrées par des preuves également lumineuses, claires et irrésistibles – car je me porte garant de la justesse de toutes ces preuves – et que la raison se voit ainsi divisée d'avec elle-même[768].

Au paragraphe 50 des *Prolégomènes à toute métaphysique future,* Kant évoque une deuxième source à la *Critique de la raison pure.* Il a recours à la même image et confie à nouveau avoir été « tiré de son sommeil dogmatique ». Mais cette fois, c'est à la lecture de David Hume qu'il attribue ce rôle déclencheur. Ainsi que le remarque Paul Clavier, le point de départ de la critique peut être l'origine du monde ou la critique humienne de la causalité, « les deux questions se recoupent dans la notion de création[769]. » Autrement dit la *Critique de la raison pure* prendrait sa source dans l'aporie à laquelle conduit la notion de création. Face à elle, la raison est dans une double impossibilité : elle semble incapable d'élucider ce qu'est la création (au sens de surgissement *ex-nihilo*), et *en même temps* incapable de renoncer à la notion.

Il est contradictoire de vouloir *élucider* l'origine d'une chose. En effet, élucider l'origine d'une chose revient à montrer en quoi cette chose était *déjà là* dans autre chose qui la précède. L'effort d'élucidation ne fait donc pas apparaître l'origine. Il ne fait que déplacer la question un cran en arrière. On dira par exemple, que l'origine de la pensée se trouve « dans » les neurones.

---

766 Emmanuel Kant, *Akademie Ausgabe,* XII, 256-258, cité et traduit par Paul Clavier, *Kant, Les idées cosmologiques,* Paris, Presses Universitaires de France, 1997, p. 78.

767 « Les antinomies n'étaient pas des supercheries, mais elles allaient au fond des choses, sous la supposition que les phénomènes et un monde sensible qui les comprend tous en lui-même seraient des choses en soi. » Emmanuel Kant, *Critique de la raison pure,* B 535, cité et traduit par Paul Clavier, *op. cit.,* p. 9.

768 Emmanuel Kant *Prolégomènes à toute métaphysique future,* § 52a, cité et traduit par Paul Clavier, *op. cit.,* p. 88.

769 *Ibid.*

Mais alors, quelle est l'origine des neurones ? Certains répondront qu'il faut regarder du côté de l'évolution. Mais l'évolution, d'où vient-elle ? Il est possible de l'attribuer à la sélection naturelle. Et la sélection naturelle ? D'aucuns diront qu'il faut étudier la vie. Mais la vie elle-même ? La régression peut s'arrêter là, si on considère que la vie est un des « trois grands mystères », mais elle peut tout autant se poursuivre : il est courant de penser qu'il serait un jour élucidé comment la vie prend sa source dans la matière. Voilà qu'on recule encore d'un cran : c'est l'origine de la matière qu'il va falloir expliquer… Là encore, quelle que soit l'origine brandie (Dieu, le Big Bang…), celle-ci ne fera que reculer d'encore un cran la question de l'origine (d'où vient Dieu ? D'où vient le Big Bang ?) – sauf à se réfugier dans un « mystère ».

Lorsqu'on lui présente l'origine de quelque chose, la raison ne se déclare jamais satisfaite : elle exige qu'on lui présente l'origine de cette origine. Ou bien on la suit et on déplace encore le problème vers l'arrière. Ou bien on s'arrête et on dit que l'origine présentée est l'origine « radicale ». Dans les deux cas, l'origine n'apparaît jamais : soit elle est renvoyée vers l'arrière, soit elle reste mystérieuse. Pour le dire en termes kantiens, il y a une « quête infatigable » de la raison vers l'inconditionné[770] qui est contradictoire avec une enquête auprès des phénomènes, puisque les phénomènes sont toujours conditionnés[771].

Aucune « origine » ne peut jamais satisfaire la raison si le terme d'origine est pris au sens strict d'un surgissement *à partir de rien*. S'il y a effectivement création (et pas seulement recombinaison d'éléments préexistants), il y a une absence avant l'apparition de la chose étudiée – un « rien » qui précède l'existence de la chose, et ce rien met l'investigation en échec puisqu'on ne peut « rien » en dire[772]. Le mécanisme n'est pertinent qu'à décrire le passage d'une chose à une autre. Il ne saurait décrire le passage de l'absence à la présence. « Rien » n'a jamais constitué une pièce valable pour une machine. Autrement dit, si la chose est apparue *de nulle part*, ce « nulle part » n'est pas un constituant valable pour une machine. Autrement dit encore, le néant ne saurait figurer dans un modèle[773].

---

770 « Mais la raison exige de connaître l'inconditionné, et avec lui la totalité de toutes les conditions, car autrement elle ne cesse de questionner tout comme s'il n'y avait pas encore eu de réponse. » Emmanuel Kant, *Akademie Ausgabe,* XX, 7, 326, cité et traduit par Paul Clavier, *op. cit.,* p. 86.

771 « C'est pour avoir voulu trouver l'inconditionné dans les phénomènes, dans les choses en tant qu'elles sont pour nous objet de l'expérience, donc toujours soumises à des conditions, que la raison métaphysicienne s'enferme dans la contradiction. » Paul Clavier, *Ibid.*

772 « Une réalité (*Wirklichkeit*) qui ferait suite à un temps vide, par conséquent une naissance (*Entstehen*) que ne précède aucun état des choses, ne peut pas plus être appréhendée que le temps vide lui-même », Emmanuel Kant, *Critique de la raison pure,* B 236-237, cité et traduit par Paul Clavier, *op. cit.,* p. 77.

773 On doit à Paul Clavier d'avoir particulièrement bien exposé cette impossibilité : « Que serait une preuve scientifique de la prémisse 'first there was nothing, then they was something ?' Est-ce que nous disposons d'un nihilomètre, d'un théoscope, ou d'un *large nothingness collider*, ou d'un autre moyen de mettre en évidence un état de chose qui exclut toute réalité physique ? Est-ce qu'on peut mettre en scène l'interaction de Dieu et du néant ? Est-ce que nous disposons d'outils théoriques qui nous permettent d'affirmer

En termes kantiens, la création ne saurait être donnée comme un phénomène, puisque cela « supprimerait l'unité de l'expérience » :

> Quand cette origine (*Ursprung aus Nichts*) est considérée comme l'action (*Wirkung*) d'une cause étrangère, on l'appelle création (*Schöpfung*) : et elle ne peut être admise comme donnée parmi les phénomènes, puisque sa seule possibilité supprimerait (*aufheben*) l'unité de l'expérience[774].

Et Paul Clavier de commenter : « l'origine radicale des choses, chère à Leibniz, n'a pas droit de cité chez Kant[775] ».

Avec la question de l'origine, la raison est comme réduite au silence. Quel que soit son choix entre les deux issues qui se présentent à elle, l'origine est soustraite à son investigation. Si elle postule un commencement absolu, ce dernier n'a pas lui-même d'origine. Si elle postule au contraire qu'il n'y a pas de commencement absolu et que le monde a toujours existé, cela revient également à évacuer la question de l'origine. Dans un cas comme dans l'autre, la raison se retrouve face à une chose *sans raison* (commencement absolu ou existence éternelle), ce qui contredit l'objet de son enquête (trouver la raison du monde). Avant d'explorer les conséquences de cette aporie pour le projet d'intelligence artificielle, nous allons brièvement exposer les quatre antinomies de la raison pure.

3.5.7. Les antinomies de la raison pure

Selon Kant, l'origine du monde n'est qu'un des différents problèmes devant lesquels la raison se trouve impuissante. Dès qu'elle considère « l'ensemble de tous les phénomènes[776] », elle se trouve face à « des *Idées,* c'est-à-dire des concepts auxquels aucun objet qui leur corresponde ne peut être donné dans l'expérience et qui, par conséquent, ne déterminent aucune connaissance véritable[777]. » Autrement dit, la raison rencontre un embarras similaire lorsqu'elle vise les choses existantes sous l'aspect de leur totalité – en tant qu'« Idées ». Dans la mesure

l'inexistence absolue, à un moment donné, de tout facteur, cause ou antécédent physique ? ». Paul Clavier, « La science est incompétente pour raisonner sur la création », *Youtube*, 10 février 2017, https://www.youtube.com/watch?v=t9BmyjR3dqM, page consultée le 20 novembre 2020

774 Emmanuel Kant, *Critique de la raison pure,* B 251, cité et traduit par Paul Clavier, *op. cit.,* p. 77.

775 Paul Clavier, *op. cit.,* p. 77.

776 Emmanuel Kant, *Critique de la raison pure,* B 446, cité et traduit par Paul Clavier, *op. cit.,* p. 7.

777 Paul Clavier, *op. cit.,* p. 7-8.

où il n'est pas possible de faire l'expérience d'une « Idée », il est impossible de vérifier ce qui est dit à leur sujet.

Kant mentionne deux autres contradictions qui surgissent lorsque le monde est considéré sous l'angle de la totalité. A l'antinomie du commencement s'ajoute l'antinomie des limites : le monde a-t-il une étendue limitée ou illimitée ? Ainsi que l'antinomie de l'atomisme : le monde est-il composé de parties simples ou bien ses parties sont-elles toujours divisibles ? De nouveau, les propositions concurrentes semblent également pertinentes, « la thèse comme l'antithèse peuvent être démontrées par des preuves également lumineuses, claires et irrésistibles » et conduisent la raison à se voir « divisée d'avec elle-même ».

Enfin, à l'« Idée » du monde s'ajoutent deux autres « Idées » que sont Dieu et l'âme, face auxquelles la raison s'égare tout autant dans sa quête de l'inconditionné. Il y donc trois « Idées » qui viennent jeter le trouble dans la raison : l'âme, le monde, et Dieu.

> Kant inventorie trois manières rationnelles – mais déraisonnables – de postuler l'inconditionné :
>
> a) Remonter à un sujet qui ne soit plus lui-même un prédicat, c'est-à-dire viser l'unité absolue du sujet pensant : c'est le propos d'une psychologie rationnelle.
>
> b) Remonter à une présuposition qui ne présuppose rien d'autre, c'est-à-dire viser l'unité absolue de la série des conditions du phénomène : projet de la cosmologie rationnelle.
>
> c) Remonter à un agrégat des membres de la division de toute la réalité, c'est-à-dire viser l'unité absolue de la condition de tous les objets de la pensée en général : ambition de la théologie rationnelle[778].

Ces trois « Idées » conduisent Kant à identifier les quatre antinomies. Les deux premières concernent le monde, tandis que les deux dernières concernent respectivement l'âme et Dieu.

La première antinomie rassemble l'aporie du commencement du monde et celle de son extension dans l'espace. Pour Kant, il s'agit du même problème – assigner une *limite du monde* – formulé dans le temps (question de l'origine) ou dans l'espace (question de l'extension)[779].

778 *Ibid,* p. 87.

779 « En effet, admettez : *Premièrement : Que le monde n'ait pas de commencement ;* il est alors *trop grand* pour votre concept ; car celui-ci, consistant dans une régression successive, ne peut jamais atteindre toute l'éternité écoulée. Posez : *qu'il y ait un commencement,* il est alors *trop petit* pour votre concept de l'entendement dans la régression empirique nécessaire. En effet, puisque le commencement présuppose toujours encore un temps qui précède, il n'est pas encore lui-même inconditionné ; la loi qui règle l'usage empirique de l'entendement vous impose de rechercher une condition de temps plus élevée encore, et par conséquent le monde est

La deuxième antinomie porte également sur le monde, mais cette fois sous l'aspect de sa composition. Ou bien le monde est composé de parties élémentaires indivisibles, ou bien la divisibilité de la matière est infinie : on peut toujours diviser une partie de la matière en parties plus simples, elles-mêmes divisibles en parties plus simples, etc[780].

La troisième antinomie vise l'Idée d'âme, et plus particulièrement l'existence de la liberté. Ou bien tout arrive selon les lois de la nature et tout est déterminé, la liberté n'existe pas ; ou bien l'âme est à l'origine de ses actions et la liberté existe. Cette dernière est alors une deuxième manière de causer les événements du monde, hétérogène aux lois de la nature[781].

La quatrième antinomie concerne l'Idée de Dieu. Ou bien il existe dans le monde un être absolument nécessaire, qui en est la cause. Ou bien aucun être nécessaire ne cause le monde, ni au sein du monde, ni en dehors[782].

Kant ne réserve pas le même traitement aux quatre antinomies. Les deux premières sont insolubles. « La thèse et l'antithèse se réfutent mutuellement[783] » sans que l'une ne l'emporte sur l'autre. Il n'est pas non plus possible de les départager en attribuant à chacune un domaine

---

manifestement trop petit pour cette loi. Il en va de même pour la double réponse faite à la question qui concerne la grandeur du monde quant à l'espace. En effet, *est-il infini* et illimité, il est alors *trop grand* pour tout concept empirique possible. *Est-il fini* et limité, alors vous demandez à bon droit : qu'est-ce qui détermine cette limite ? L'espace vide n'est pas un corrélat des choses qui subsiste en lui-même, et ne peut être une condition empirique constituant une partie d'une expérience possible. (Car qui peut avoir une expérience du vide absolu ?) Mais pour l'absolue totalité de la synthèse empirique, il est toujours exigé que l'inconditionné soit un concept d'expérience. Un *monde limité* est donc *trop petit* pour votre concept. » Emannuel Kant, *Critique de la raison pure,* B 514-517, cité et traduit par Paul Clavier, *op. cit.*, p. 93.

780 « *Deuxièmement :* Si tout phénomène dans l'espace (toute matière) se compose *d'une multiplicité infinie de parties (aus unendlich viel Teilen),* la régression de la division est toujours *trop grande* pour votre concept ; et si la *division* de l'espace doit *s'arrêter* à quelqu'un de ses membres (au simple), cette régression est *trop petite* pour l'idée de l'inconditionné. En effet ce membre laisse toujours encore de la place pour une régression vers un plus grand nombre de parties contenues en lui. » *Ibid*, p. 94.

781 « *Troisièmement :* Si vous admettez qu'en tout ce qui arrive dans le monde, il n'y rien qui ne soit conséquence (E*rfolg)* selon des lois de la nature, alors la causalité de la cause *(Kausalität der Ursache)* est toujours à son tour quelque chose qui arrive, et elle vous force sans cesse à remonter dans votre régression à des causes plus élevées encore, et par conséquent à prolonger sans arrêt la série des conditions *a parte priori.* La simple *nature* productrice d'effets *(wirkende)* est donc trop grande pour tout votre concept, dans la synthèse des évenements du monde *(Weltbegenbenheiten)*. Si vous optez, ça et là, pour des événements produits *d'eux-mêmes (von selbst gewirkte Begenbenheiten),* par conséquent une production *(Erzeugung)* par liberté, la question du pourquoi, selon une inévitable loi de nature, vous poursuit et vous force à remonter au-delà de ce point suivant la loi causale de l'expérience, et vous trouverez alors qu'une semblable totalité de la connexion *(Verknüpfung)* est trop petite pour votre concept empirique nécessaire. » *Ibid.*

782 « *Quatrièmement :* Si vous admettez un être absolument (*schlechthin)* nécessaire (que ce soit le monde lui-même, ou quelque chose dans le monde, ou la cause du monde), vous le placez dans un temps infiniment éloigné de tout instant (*Zeitpunkt)* donc, puisque autrement il serait dépendant d'une autre existence plus ancienne ; mais alors cette existence est inaccessible *(unzugänglich)* à votre concept empirique, et elle est *trop grande* pour que vous puissiez jamais y arriver par quelque régression continue. Mais si, dans votre perspective, tout ce qui appartient au monde (que ce soit comme conditionné ou comme condition) est *contingent,* alors toute existence qui vous est donnée est *trop petite* pour votre concept. En effet, elle vous force à chercher toujours encore une autre existence d'où elle dépende. » *Ibid*, p. 94-95.

783 Paul Clavier, *op. cit.*, p. 96.

différent car elles ont « affaire à l'addition ou à la division de ce qui est homogène (espace, temps et matière considérés comme des grandeurs[784]) ».

Par contre, il est possible de trouver une issue aux deux dernières antinomies en attribuant une portée différente à chaque proposition : la thèse vise les choses en soi, tandis que l'antithèse concerne les phénomènes. Ainsi, les deux propositions peuvent être vraies en même temps et ne se contredisent plus puisqu'elles ne parlent pas des mêmes choses[785].

En ce qui concerne la troisième antinomie ou antinomie de la liberté, cela revient à considérer le sujet sous deux aspects :

> […] un *caractère empirique,* par lequel ses actions, comme phénomènes, seraient totalement prises dans l'enchaînement avec d'autres phénomènes suivant des lois constantes de la nature […], on devrait accorder au sujet, en outre, un *caractère intelligible,* par lequel à la vérité il est la cause de ses actes comme phénomènes, mais qui lui-même n'est soumis à aucune des conditions de la sensibilité et n'est pas même un phénomène[786].

Si Kant réussit à tirer la raison de l'embarras, il en vient tout de même à une conclusion négative. Il est possible d'étudier le « caractère empirique » du sujet, mais pas son « caractère intelligible ». Ce dernier ne saurait faire l'objet d'une expérience puisqu'il « n'est pas même un phénomène ». Cela condamne toute perspective d'une science de la liberté. Comment pourrait-on étudier le choix puisque l'étude ne fera voir que les causalités à l'œuvre, et jamais le choix lui-même ?

3.5.8. Paralogismes de la raison et « défis » de l'IA

En cherchant à étudier le « caractère intelligible » du sujet, le projet d'intelligence artificielle ignore la distinction entre phénomène et noumène et s'empêtre dans des contradictions

784 *Ibid.*

785 Voici comment Kant le formule au sujet de la quatrième antinomie : « puisque les deux thèses en conflit peuvent être vraies en même temps sous des rapports différents, de telle sorte que toutes les choses du monde sensible soient entièrement contingentes et par conséquent n'aient toujours aussi qu'une existence empiriquement conditionnée, et qu'il y ait pourtant aussi pour toute la série une condition non empirique, c'est-à-dire un être inconditionnellement nécessaire. Celui-ci en effet, en tant que condition intelligible, n'appartiendrait pas du tout à la série comme un de ses membres (pas même comme son membre le plus élevée)… », Emmanuel Kant, *Critique de la raison pure,* B 588, cité et traduit par Paul Clavier, *op. cit.*, p. 97-98.

786 Emmanuel Kant, *Critique de la raison pure,* B567, traduit et cité par Paul Clavier, *op. cit.*, p. 97.

d'apparence insolubles, tout comme on s'empêtrerait dans les antinomies de la raison pure si l'on voulait recommencer un monde. De la même manière que la raison ne peut trancher entre l'idée que le monde a un commencement et celle qu'il n'en a pas, il est impossible de trancher au sujet de l'émergence de l'intelligence : ou bien elle a surgi *ex-nihilo* et on reste impuissants à en rendre compte, ou bien elle a émergé de la matière et c'est l'émergence d'une matière contenant l'intelligence qu'il faut expliquer. Ce qui vaut pour l'intelligence vaut pour les idées individuelles : ou bien nous sommes à l'origine de nos idées (c'est le point de vue de Lovelace) et ce qui nous arrive est inexplicable, ou bien c'est explicable et nous ne sommes à l'origine de rien (c'est le point de vue de Turing en 1950).

De la même manière, le projet d'intelligence artificielle se heurte à la troisième antinomie de la raison pure : vouloir fabriquer un programme qui soit « à l'origine » de ses idées, c'est-à-dire une machine « autonome » au sens où elle est à l'initiative de ses décisions, c'est vouloir conditionner un sujet, autrement dire conditionner l'inconditionné. Il s'agit donc de faire tenir ensemble les deux propositions incompatibles de la troisième antinomie : la notion de programme répond à l'idée que « tout a une cause et tout est déterminé selon les lois de la nature » et la notion de programme libre reprend celle que « le sujet libre est une exception à cette détermination, il se cause lui-même ». En d'autres termes, fabriquer une machine autonome est aussi contradictoire que de vouloir forcer quelqu'un à être libre.

On doit à Jerry Fodor d'avoir eu l'honnêteté intellectuelle de pousser si loin son analyse de l'hypothèse de la modularité de l'esprit qu'il en est arrivé à mettre en lumière son caractère contradictoire, qui ressemble à celui de la deuxième antinomie (le monde est composé de parties simples / le monde est composé de parties elles-mêmes composées). L'étude des « modules » de l'esprit l'a amené à considérer comment certaines fonctions (la coordination par un « système central », la conscience, l'invention…) ne peuvent être modulaires puisqu'elles doivent pouvoir s'appliquer à chacun des modules indistinctement, sans que l'on puisse en définir les règles d'application[787].

Lorsque la raison s'observe avec les lunettes de la science (sous le nom d'intelligence), elle se conçoit comme une chose fabriquée, décomposable en parties simples et sujette à la causalité ; mais pour peu qu'elle se considère avec les yeux de l'expérience (sous le nom d'intuition), elle se conçoit comme une chose spontanée, « générale » et libre, ou au moins « créative ». Les deux points de vue entrant en contradiction, l'origine de la pensée, son unité et sa liberté semblent constituer autant de domaines où la science est impuissante et autant

787 Jerry Fodor, *La modularité de l'esprit : essai sur la psychologie des facultés,* Paris, Les Éditions de Minuit, 1986.

d'objections au projet d'intelligence artificielle : comment la pensée pourrait-elle être à la fois fabriquée et spontanée, modulaire et générale, déterminée et libre ?

Plutôt que d'en conclure à l'impossibilité du projet d'intelligence artificielle, les chercheurs préfèrent ignorer les apories de la genèse, de la composition, et de la liberté de la raison en se concentrant sur les « défis » techniques que représentent « l'émergence », la « modularité » et la « créativité » de l'intelligence. Les apories ne sont pas perçues commes des limites, mais au contraire comme l'horizon du projet d'IA, c'est le but à accomplir. Les contradictions du mécanisme ne remettent pas leur élan en question. Au contraire, ils n'y voient que l'invitation à fabriquer de meilleures machines.

3.5.9. La séduction par les Idées

Dans une formule qui résonne singulièrement avec les « big magnificent questions » évoquées par Feigenbaum ou les « questions scientifiques qui résument la science » de LeCun, Kant souligne l'attrait irrésistible des Idées : les quatre antinomies de la raison pure soulèvent « des questions pour la solution desquelles le mathématicien donnerait volontiers toute sa science[788]. » C'est la perspective de l'inconditionné qui séduit la raison jusqu'à l'éloigner de l'expérience (du conditionné) et lui faire perdre pied. Kant compare la métaphysique en quête de réponses définitives à un navigateur en quête d'une terre ferme. L'un comme l'autre ne cessent de se leurrer en prenant pour des solutions ce qui ne sont que des mirages – « maint banc de brume, mainte banquise sur le point de fondre se présentent trompeusement comme de nouveaux pays et ne cessent de l'abuser par de vaines espérances[789] ». Bien qu'il semble vain de poursuivre les recherches, le navigateur se refuse d'abandonner l'espoir de trouver une terre ferme. Le voilà donc, et le philosophe métaphysique avec lui, « empêtré dans des aventures, auxquelles il ne peut renoncer, mais qu'il ne peut jamais conduire à bonne fin[790] ».

Kant lui-même s'est trouvé empêtré dans de telles aventures. Sa fascination pour les travaux de Newton et les lois de la gravitation l'ont amené à diverses spéculations cosmologiques[791]. Étrangement, celles-ci ne seront pas abandonnées après la rédaction de la

788 Emmanuel Kant, *Critique de la raison pure,* B490, traduit et cité par Paul Clavier, *op. cit.*, p. 91.
789 Emmanuel Kant, *Critique de la raison pure,* B295, traduit et cité par Paul Clavier, *op. cit.*, p. 80.
790 *Ibid.*
791 Paul Clavier évoque « l'enthousiasme cosmologique de Kant » dans son « projet d'une synthèse de la métaphysique wolffienne avec la mécanique newtonienne, c'est-à-dire la combinaison d'une doctrine dans

Critique de la raison pure. Il y a là un « intérêt persistant […] pour la spéculation cosmologique, intérêt jamais démenti malgré la censure dont l'*Antinomie de la raison pure* semblait avoir frappé les idées cosmologiques[792] ». A titre d'exemple, Paul Clavier mentionne certains écrits ultérieurs à la *Critique de la raison pure* où Kant entre en contradiction flagrante avec ses propres propos.

> Dans tel autre moment d'enthousiasme cosmologique (*Reflexionen,* n° 4966), Kant proclame résolue la difficulté de la métaphysique : la simple idée cosmologique est empiriquement réalisée dans l'architecture gravitationnelle de la matière. Évidemment, une telle solution fait bon marché de l'idéalisme transcendental, solution officielle de la dialectique cosmologique. C'est ce que Kant reconnaît expressément dans une liasse de l'*opus postumum*[793] […].

Kant aurait-il passé outre ses propres avertissements et se serait-il contredit ? Paul Clavier préfère en rester au caractère marginal des passages incriminés :

> On dira que ces déclarations, que Kant multiplie dans l'*opus postumum,* restent marginales par rapport à l'œuvre publiée, et que Kant prêche ici le faux pour savoir le vrai, ou bien que quelque mouche l'a piqué[794].

Une autre interprétation est possible. Si Kant a si bien évoqué l'irresistible attraction de l'inconditionné pour la raison, c'est peut-être parce qu'il n'a cessé lui-même de la ressentir. La persistance des spéculations cosmologiques au long de son œuvre illustrerait son propos au lieu de le contredire. Tout en sachant que ces spéculations sont vouées à la désillusion, Kant n'aurait pu s'empêcher, comme le navigateur, de cultiver l'espoir d'une terre ferme. La spéculation sur la gravitation universelle est « une de ces aventures, auxquelles il ne peut renoncer ». Le projet d'intelligence artificielle participe-t-il de ces errances métaphysiques, s'agit-il d'une nouvelle aventure de la raison, à laquelle elle ne peut renoncer, tout en étant incapable de la faire aboutir ?

---

laquelle l'univers résulte d'une connexion d'éléments, avec la représentation du système du monde au moyen des lois mathématiques, notamment la loi de pesanteur universelle. », *op. cit.,* p. 106.

792 Paul Clavier, *op. cit.,* p. 104.

793 *Ibid,* p. 112.

794 *Ibid,* p. 113.

3.5.10. L'erreur prosyllogistique

C'est l'attrait de l'absolu, la séduction qu'exerce l'inconditionné, qui amène la raison loin de l'expérience, jusqu'à tenir des propositions contradictoires. En prenant pour horizon « l'intelligence générale » ou le « principe » général gouvernant l'ensemble des « modules particuliers » de l'intelligence, la raison se laisse aller aux mêmes errements que Kant décrit à propos des antinomies. Grisés par notre capacité à étudier chaque phénomène qui se présente, nous sommes amenés à croire que *la totalité* des phénomènes est également un phénomène (donc observable), et que dans celle-ci se trouve *le principe* commun à chaque phénomène. Nous adhérons au raisonnement suivant, que Kant dénonce comme fallacieux : « si le conditionné est donné, se trouve donnée aussi la somme entière des conditions, et par conséquent l'absolument inconditionné, par lequel seulement le conditionné était possible[795] ». Suite à l'étude fructueuse de conditions données, la raison conclut, trop vite, qu'est également possible l'étude de l'ensemble des phénomènes, lui-même « absolument inconditionné ». Ainsi la raison est-elle amenée à penser que l'inconditionné peut s'observer, que l'absolu est à portée de lunette. En l'occurence, déduire de la possibilité d'étudier les manifestations de l'intelligence qui se présentent à nous, qu'il est possible d'observer l'origine et la condition de tous les objets de pensée en général.

Kant appelle « prosyllogisme » l'erreur consistant à remonter des conditions vers un principe, puis à ériger ce principe comme une chose. Puisque *chaque* condition est donnée, on en déduit que *toutes* les conditions sont données, et ensuite que la *totalité* des conditions est aussi un objet qui se donne. En ce qui concerne le projet d'intelligence artificielle, cela revient à adopter le point de vue suivant : puisque *chaque* aspect de l'intelligence peut être étudié, alors l'intelligence *en tant que totalité* peut également être étudiée. L'intelligence, en tant que totalité des aspects de l'intelligence, ou comme principe unificateur, est prise comme un objet de même ordre que les aspects particuliers. Il est postulé « qu'au-dessus » de l'ensemble des conditions particulières (en tant que totalité) ou « en-dessous » (en tant que principe générateur), existe un objet aussi objectif que les conditions particulières, mais inconditionné car lui-même à l'origine des conditions particulières. La raison postule l'existence d'une unité : l'intelligence comme principe.

795 Emmanuel Kant, *Critique de la raison pure*, traduction Renaut, Paris, GF Flammarion, 2001, p. 419.

Le problème ne vient pas de l'identification de principes unifiant la diversité des phénomènes – parfois désignés sous le terme de « lois de la nature ». Il s'agit là du mouvement même de la science. Ce qui distingue le travail de la science de l'erreur de la régression prosyllogistique dénoncée par Kant a lieu lorsque le principe est pris comme une chose. En droit, chaque phénomène peut être étudié, et ses conditions élucidées. Mais cela ne permet pas de préjuger de la totalité des phénomènes car aucune expérience ne peut valider de proposition à son sujet.

3.5.11. Le logiciel de l'âme

Entreprendre la fabrication d'une machine pensante, c'est postuler la possibilité de matérialiser *ce qui pense,* autrement dit le sujet, ici compris comme la source de chaque manifestation d'intelligence, ou encore le principe sous-jacent (la sub-stance) de chaque pensée. En s'efforçant d'atteindre la condition de toutes les pensées, le point inconditionné où converge la regression des phénomènes vers leurs causes, le projet d'intelligence artificielle correspond précisément à un cas d'erreur prosyllogistique. Viser « l'unité absolue du sujet pensant » est une des trois manières déraisonnables de postuler l'inconditionné. Séduite par une des trois Idées – ici l'Idée d'âme –, la raison se laisse entraîner dans une quête de l'inconditionné qui l'amène à s'égarer loin des limites de l'expérience.

Le projet d'intelligence artificielle tel qu'il est formulé à Dartmouth se contente de postuler que « tous les aspects de l'apprentissage ou toute autre caractéristique de l'intelligence peuvent en principe être décrits » mais il n'y a qu'un pas à franchir pour aller de *toutes* les caractéristiques de l'intelligence à l'intelligence comme *totalité de ces caractéristiques* ou plus précisément comme *principe* régissant la totalité de ces caractéristiques. La tentation est grande de passer de la diversité des conditions (les différentes caractéristiques de l'intelligence) à l'unité d'un principe (l'intelligence comme facteur commun aux différentes facultés, ou « facteur g »). Tout comme Kant a cherché dans la gravitation universelle « la découverte d'un fondement positif matériel qui réaliserait empiriquement l'idée cosmologique[796] », le projet d'intelligence artificielle aspire à découvrir le principe « qui réaliserait empiriquement » l'idée d'un sujet libre.

796 *Ibid*, p. 105.

L'Idée séductrice en cause est celle d'âme, définie par Kant comme le postulat d'une unité absolue du sujet pensant. Pour Kant, le but de la critique n'est pas de se débarrasser de l'Idée d'âme, mais au contraire de lui faire une place en tant qu' « Idée régulatrice », ce qui aura toute son importance au moment de la *Critique de la raison pratique.* Par contre, il condamne sans appel tout projet de science de l'âme. Il est possible d'étudier les manifestations de la pensée (c'est la « psychologie empirique », selon les termes de Wolff) mais c'est se leurrer que de vouloir remonter de celles-ci à un principe (« psychologie rationnelle »). L'Idée d'âme peut et doit guider la raison pratique, mais elle ne saurait faire l'objet d'une science. Il peut y avoir une science du sujet, une « psychologie », sur le modèle des sciences physiques, mais, comme pour ces dernières, la science doit en rester à l'expérience et se garder de toute considération sur l'unité ou la totalité des phénomènes. Il y a une constance dans la façon dont je me désigne « moi », mais cela n'entraîne pas qu'il y ait effectivement une constante sous-jacente (un « moi » immuable) à toutes ces désignations. Sur ce point, Kant reprend et adapte la critique de Hume[797]. L'erreur consiste à passer du « je pense », sujet transcendental présupposé par l'unité de mes perceptions (sujet empirique) à un « je pense » comme substance sous-jacente à toutes les expériences subjectives (sujet transcendant). L'existence du sujet comme inconditionné ne vaut que subjectivement, rien ne permet de le poser comme une réalité en soi et d'en faire un objet de science.

Du côté des informaticiens, mathématiciens, physiciens, psychologues, neuroscientifiques et philosophes qui se réclament de la recherche en intelligence artificielle, le terme d'âme est généralement évité, à cause de ses connotations religieuses. Au lieu du vieux mot d'âme, ils emploient ceux d'esprit (*mind*), de conscience (*consciousness*) ou, encore plus prosaïque, celui d'intelligence. Mais si les mots sont différents, l'appareil conceptuel est similaire : celui d'une distinction entre la matière et un principe immatériel qui l'anime – le matériel (*hardware*) et le logiciel (*software*) – et la recherche de ce principe immatériel. Le mot « âme » n'est pas mentionné, mais c'est bien le principe immatériel animant la matière qui est l'objet de l'enquête. Conformément à la tradition sémitique, ce principe immatériel est un « code » dans les deux sens du terme : un texte caché et un ensemble de règles qui commande l'agencement de la matière. Dans la mesure où est recherché *ce qui fait que la matière pense*

797 « Pour moi, quand je pénètre le plus intimement dans ce que j'appelle *moi-même,* je tombe toujours sur une perception particulière ou sur une autre, de chaleur ou de froid, de lumière ou d'ombre, d'amour ou de haine, de douleur ou de plaisir. Je ne parviens jamais, à aucun moment, à me saisir *moi-même* sans une perception et je ne peux jamais rien observer d'autre que la perception. » David Hume, *L'entendement, Traité de la nature humaine, livre I et Appendice*, traduction de Philippe Baranger et Philippe Saltel, Paris, GF Flammarion, 1995, p. 343. La section IV développe ce thème « De l'identité personnelle », p. 342-355.

dans une suite d'instructions, alors cette suite d'instructions, même en ayant la prudence d'en rester aux termes de « logiciel », d'esprit, de conscience ou d'intelligence, est bien un recyclage de l'Idée d'âme. Il y a bien une réification de l'âme puisque c'est une *chose* qui est recherchée – le logiciel.

Dès l'invention des ordinateurs programmables[798], un certain nombre d'orientations techniques ont favorisé l'analogie entre l'informatique et le dualisme corps-esprit : distinction entre matériel et logiciel, passage de l'analogique au numérique (le signal numérisé, plus facile à reproduire, s'émancipe de son support) et prééminance d'une théorie de l'information fondée sur le code plutôt que le signal qui conçoit l'information comme indépendante du medium[799]. Comme l'a souligné Katherine Hayles, à mesure que l'information s'autonomise du medium et en vient à passer pour un « fluide » immatériel, désincarné, capable de circuler de support en support sans modification[800], c'est notre propre dualisme corps-esprit qui s'en trouve renforcé. L'informatique passe pour une imitation de la séparation corps-esprit et semble confirmer que tout se joue au niveau de l'esprit. Dès lors, il n'y a qu'un pas à faire, que Hans Moravec fera avec *Mind Children,* pour affirmer que l'« identité humaine » ou la conscience, n'est que de l'information, et qu'à l'instar de celle-ci, elle peut s'émanciper de son support – le corps[801]. Le mot « âme » n'est jamais utilisé mais ce dont on parle, conçu comme immatériel et incorruptible, y ressemble fortement.

Le langage courant décrit la circulation de l'information dans le réseau comme une circulation du même, oblitérant ainsi l'immense travail de duplication, de reconstruction, de vérification de l'information qui permet au final de la faire apparaître comme « identique ». Ainsi on apprendra qu'un « fichier est corrompu » (*file is corrupt*) et non qu'il y a eu échec dans les opérations de duplication, reconstruction et vérification (*file failed to rebuild*). En réalité, l'« identité » de l'information tient plus de la recréation permanente que de la circulation

798 On parle d'ordinateur programmable lorsque le même ordinateur peut exécuter plusieurs programmes distincts.

799 « Chez Wiener, l'information sera toujours regardée à la manière d'un signal comme l'expression d'un certain ordre matériel, alors que chez Shannon, l'information participe de l'ontologie fantomatique du signe. » Mathieu Triclot, *Le moment cybernétique : La constitution de la notion d'information,* Seyssel, Éditions Champ Vallon, 2008, p. 26.

800 « Aiding this process was a definition of information, formalized by Claude Shannon and Norbert Wiener, that conceptualized information as an entity distinct from the substrates carrying it. From this formulation, it was a small step to think of information as a kind of bodiless fluid that could flow between different substrates without loss of meaning or form. », Katherine Hayle, *How We Became Posthuman : Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, University of Chicago Press, 2008, p. xi. À la différence Hayles, Mathieu Triclot voit une différence entre la notion d'information selon Wiener et selon Shannon.

801 « Writing nearly four decades after Turing, Hans Moravec proposed that human identity is essentially an informational pattern rather than an embodied enaction. The proposition can be demonstrated, he suggested, by downloading human consciousness into a computer, and he imagined a scenario designed to show that this was in principle possible. » *Ibid,* p. xi – xii.

du même, mais l'oubli du travail actif de préservation et de maintenance permet de faire apparaître l'information et le logiciel comme une chose immatérielle qui aspire à circuler « librement », c'est-à-dire sans altération, sans *corruption,* indemne (non *damnée*). Le travail de maintenance est gommé et n'apparaît que quand la circulation échoue. Chaque panne, au lieu de rappeler que tout système ne tient que grâce à un travail constant, est au contraire l'occasion de blamer « l'erreur humaine » ou la réticence de la matière, qui viendraient entraver la circulation idéale de l'information. Alors que les humains et les machines copient et préservent activement l'information, leur matérialité crasse est accusée de faire obstacle à sa libre circulation. Comme si, à cause de quelques frottements ou de fuites, les canalisations étaient accusées d'empêcher la circulation du gaz. Rolf Landauer avait bien vu ce qui se tramait de théologique dans les théories de « l'information libre » et remarquait qu'« il n'y a pas de distinction aux relents spiritualistes entre l'information libre et l'information liée. Il n'y a que de l'information liée[802]. » L'information ne peut être qu'un processus physique. Mathieu Triclot commente : « Le *software* n'existe pas. L'arrière-monde des symboles et du traitement algorithmique pur de l'information doit être ramené du ciel sur la terre[803]. » Mais ce n'est pas cette théorie de l'information qui a prévalu. Selon les termes de Mathieu Triclot, le « code » l'emporte sur le « signal », permettant d'en faire le support d'une analogie avec l'âme, d'imiter le dualisme corps-esprit dans les dualismes information-support et logiciel-matériel, et de proposer une forme de métempsychose via la libre circulation de l'information et le principe d'implémentation multiple :

> Une même fonction peut par exemple être implémentée sur un ordinateur ou un cerveau, ou bien sur un processeur de silicium ou un 'processeur' en fromages, chats et souris. Cette position implique deux principes : la matière est indifférente à la fonction (dégagement d'un niveau d'analyse autonome des processus psychiques) et une fonction peut être réalisée par des dispositifs tout à fait différents (principe d'implémentation multiple)[804].

Si l'âme est le logiciel du corps, fabriquer une machine qui pense reviendrait à mettre au point le logiciel capable d'animer la matière comme notre âme anime notre corps. En perçant les secrets de l'âme, l'intelligence artificielle fait miroiter la perspective de « téléchargements de l'esprit » et remet au goût du jour les rêves d'immortalité et de métempsychose. Si les

802 Ralf Landauer, « Computation : a fundamental physical view », *Phys. Scr.*, 35, 1987, p. 88-95, in Harvey Leff et Andrew Rex (dir.), *Maxwell's Demon, Entropy, Information, Computing,* Princeton, Princeton University Press, 1990, p. 260-262. cité par Mathieu Triclot, *op. cit.*, p. 274.
803 *Ibid.*
804 *Ibid*, p. 149.

chercheurs contemporains comme LeCun, Bengio ou Hinton se gardent bien de déclarations à ce sujet, c'est une interprétation courante du projet d'intelligence artificielle, abondamment représentée par les ouvrages de science-fiction[805], à laquelle adhèrent sans réserve les courants transhumanistes[806], et dont le chantre le plus éloquent est probablement Ray Kurzweil[807].

3.5.12. La lecture éliminativiste

Avec le paradigme connexionniste, le projet d'intelligence artificielle semble amorcer un retour vers le corps et se détourner d'un tel projet. Le principe de l'intelligence n'est plus recherché dans une « théorie de l'esprit » mais dans la bonne manière d'agencer les neurones, un agencement qui se produit par « apprentissage », c'est-à-dire en fonction des données qu'on lui fournit, et dans des tâches plus près du corps (vision…) que de l'esprit. Nous l'avons vu, cela apparaît comme un retour à la cybernétique et au paradigme du signal, c'est-à-dire à un ancrage dans la matérialité de l'information. En apparence, les chercheurs connexionnistes s'opposent donc à l'analogie entre l'âme et le logiciel : la fabrication d'une machine pensante ne percera pas les secrets de l'âme mais prouvera au contraire que la matière n'a pas besoin d'être « animée » pour penser. Orienter la recherche sur la notion d'intelligence plutôt que sur celle du sujet ou de l'âme c'est sous-entendre que *ce qui pense* n'est pas le sujet mais un principe non-subjectif, une fonction – l'intelligence. Ils pensent démontrer qu'il peut y avoir de la pensée sans sujet, quitte à en conclure que les humains en sont peut-être tout aussi dépourvus. Ils n'ont pas plus besoin d'âme ou de sujet pour penser que les machines. Le secret de l'âme, c'est qu'elle n'existe pas[808]. Conformément au cliché du scientifique « chasseur de mythe[809] » et prolongeant le « désenchantement du monde[810] », le projet d'intelligence artificielle viendrait s'attaquer ici à une des dernières superstitions du monde moderne.

---

805 Par exemple la série *Upload*, certains épisodes de *Black Mirror* (*San Junipero, Be right back), Transcendance* ou *Ghost in the Shell.*

806 Voir l'enquête du journaliste Mark O'Connell auprès de différents groupes transhumanistes : *To Be A Machine : Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death,* Londres, Granta Books, 2017.

807 Ray Kurzweil, *The Age of Spiritual Machines,* New-York, Viking Press, 1999.

808 C'est la lecture dite « éliminativiste » de la philosophie de l'esprit. Michael Esfeld, *La philosophie de l'esprit : Une introduction aux débats contemporains, op. cit.*, chapitre 12, section 4.

809 L'expression célèbre est de Norbert Elias, au sujet des sociologues. Norbert Elias, *Qu'est-ce que la sociologie,* Paris, Pocket, 1993.

810 Thème célèbre de l'œuvre de Max Weber. Le processus de rationalisation, déjà à l'œuvre dans l'évolution des religions (passage du catholicisme au protestantisme), est prolongé par la place croissante prise par les savoirs

Cette lecture du projet d'intelligence artificielle, qui pourrait être qualifiée d'« éliminativiste », n'échappe pas à la critique. Pour Kant, il y a autant de dogmatisme chez ceux qui nient l'existence de l'âme que chez ceux qui l'affirment. En voulant prouver que l'âme n'existe pas, la raison outrepasse autant les limites que lorsqu'elle cherche à prouver qu'elle existe. Dans les deux cas, il s'agit de mettre le doigt sur ce qu'elle est – quitte à montrer qu'elle n'est rien. Or l'Idée d'âme ne peut donner lieu qu'à des spéculations, et non à une expérience permettant de trancher définitivement. Ainsi, matérialistes et spiritualistes sont renvoyés dos à dos et accusés de la même erreur : prendre l'existence ou la non-existence comme une propriété du concept et se laisser entraîner à des hypothèses que l'expérience ne peut vérifier. Vouloir montrer que *rien* ne pense en faisant advenir une pensée *sans sujet* participe aussi de l'« erreur prosyllogistique », de cette forme d'errance de la raison que Kant condamne. Que l'on postule l'existence ou l'*inexistence* d'une substance comme substrat à la pensée, on tombe dans l'erreur qu'il dénonce. Plus généralement, la pespective éliminativiste souffre de l'erreur consistant à méconnaître que l'invalidation des discours métaphysiques est *déjà* un discours métaphysique. Dire que « la métaphysique n'existe pas », c'est faire de la métaphysique.

Tableau récapitulatif

La raison a des limites, notamment en ce qui concerne son propre fonctionnement / Il y a des questions sans réponses.	La raison peut élucider tous les problèmes, et donc son propre fonctionnement / Toutes les questions ont une réponse.
« Ignoramus et ignorabimus » (Emil du Bois-Reymond)	« Nous devons savoir et nous saurons » (Hilbert)
Il existe des aspects de l'apprentissage ou de l'intelligence qui ne peuvent être décrits d'une manière si précise qu'une machine peut les simuler (recherche d'un « résultat négatif » invalidant la conjecture de Darmtouth).	« […] tous les aspects de l'apprentissage ou toute autre caractérisitique de l'intelligence peuvent en principe être décrits d'une manière si précise qu'une machine peut les simuler » (conjecture de Dartmouth).
L'intuition (en tant que mode de pensée non mécanique) existe (Turing 1938, Gödel…)	L'intuition n'existe pas ou bien l'intuition est un mode de pensée mécanique non conscient (Turing 1950, GOFAI, connexionnisme).

scientifiques qui viennent remplacer les anciennes superstitions. Max Weber, *L'éthique protestante et l'esprit du capitalisme*, Paris, Flammarion, 2017.

L'intuition peut se décrire après-coup mais elle ne s'explique pas. Elle est comme un « effet sans cause ».	Le fonctionnement de l'intuition peut s'expliquer au même titre que n'importe quel autre phénomène.
Les phénomènes ne peuvent nous apparaître que selon un ordre, mais nous ignorons ce qu'il en est du monde en soi.	Le monde est ordonné (pythagorisme).
La science décrit les phénomènes.	La science décrit le réel (réalisme scientifique).
La science est impuissante à rendre compte du devenir.	La science peut rendre compte du devenir.
Le devenir est sans loi (évolutionnisme véritable, pas de loi de l'évolution).	Le devenir est régi par des lois.
Les mythes permettent de parler de ce qui échappe à la raison (notamment les questions de création / origine / nouveauté / liberté / devenir).	Les mythes doivent être remplacés par la science (notamment l'émergence du cosmos, de la vie et de la conscience).
L'intelligence artificielle forte est impossible (mais les machines peuvent faire illusion).	L'intelligence artificielle forte est possible.

3.6. Les raisons de la raison

3.6.1. L'intuition est sans pourquoi

Dans la préface de *Sylvie et Bruno*, Lewis Carroll rapporte que les éléments du livre lui sont venus en une suite désordonnée d'idées improbables (« all sorts of odd ideas ») à des moments tout aussi improbables (« at odd moments »). S'il a pu retracer l'origine de certaines de ces fulgurances (« random flashes of thought ») dans une lecture ou la suggestion d'un ami, elles sont aussi venues d'elles-mêmes, sans raison (« a propos of nothing »), comme autant de « spécimens de ce phénomène désespérement illogique » (« specimens of that hopelessly illogical phenomenon »), comme un « effet sans cause » (« an effect without a cause[811] »). Ce fut le cas, par exemple, de la dernière ligne de 'The Hunting of the Snark', qui lui est venue pendant une marche solitaire, et de quelques passages qui lui sont venus en rêve, sans qu'il puisse les rapporter à aucune cause (« which I cannot trace to any antecedent cause whatever[812] »). Avec l'intuition, nous faisons l'expérience d'un effet sans cause, sans savoir quelle est la validité de cette expérience. S'agit-il d'une illusion ? Ou bien l'effet est sans cause et nous assistons au surgissement des idées *ex-nihilo ;* ou bien l'effet a une cause qui nous est cachée. L'introspection ne livre aucune explication. Pour les tenants de l'intelligence artificielle, cela ne permet pas de conclure qu'il n'y a pas de cause : le fonctionnement de l'intuition peut s'expliquer au même titre que n'importe quel autre phénomène. Ce n'est pas que la cause manque, c'est que nous l'ignorons. Ils adhèrent au principe de raison et récusent qu'il y ait le moindre surgissement *ex-nihilo*. Pour eux, l'émergence d'une idée est un processus neurophysiologique. Elle peut, comme l'émergence de la vie ou de la conscience, se réduire à un certain agencement de matière. Le principe de raison est suivi scrupuleusement. Pour le

811 « As the years went on, I jotted down, at odd moments, all sorts of odd ideas, and fragments of dialogue, that occured to me – who knows how ? – with a transitory suddenness that left me no choice but either to record them then and there, or to abandon them to oblivion. Sometimes one could trace to their source these random flashes of thought – as being suggested by the book one was reading, or struck out from the 'flint' of one's own mind by the 'steel' of a friend's chance remark but they had also a way of their own, of occurring, a propos of nothing – specimens of that hopelessly illogical phenomenon, 'an effect without a cause'. » Lewis Carroll, *Sylvie and Bruno,* Londres, Macmillan and Co., 1889.

812 « Such, for example, was the last line of 'The Hunting of the Snark', which came into my head (as I have already related in 'The Theatre' for April, 1887) quite suddenly, during a solitary walk : and such, again, have been passages which occurred in dreams, and which I cannot trace to any antecedent cause whatever. » *Ibid.*

préserver, les surgissements seront présentés comme des illusions : la vie ou la conscience ne sont que des façons d'organiser la matière de manière à ce qu'elle puisse « passer pour » vivante ou consciente – c'est ce que laisse entendre le test proposé par Turing, il vaut mieux se demander si une entité peut « passer pour » pensante que de chercher à savoir si elle pense vraiment.

Sans pour autant sombrer dans la confusion, il est possible de desserrer l'étreinte du principe de raison pour mieux appréhender la question de l'origine. Nous avons vu que la régression des causes finit par heurter un inconditionné. Si la cause de la pensée est dans la matière, quelle est la cause de la matière ? Quelle que soit la raison invoquée – Dieu ou une matière ayant toujours existé – le problème est encore déplacé sans être résolu. Ou bien cette chose a elle-même une cause ou bien, si elle est l'origine, elle n'a pas de cause, elle est sans raison[813]. Autrement dit, à force de remonter la chaîne des causes à la recherche de l'origine, la raison ne peut qu'arriver à une origine *sans cause*. Penser qu'il serait possible d'éviter le problème en affirmant qu'*il n'y a pas d'origine* revient à retomber dans le même silence de la pensée, ou plutôt à l'admettre encore plus vite : l'impossibilité à penser l'origine est équivalente à l'impossibilité de penser l'absence d'origine. Renvoyer, comme le fait Yann LeCun, à une explication à venir des « mystères » de la conscience, la vie, la matière, n'est qu'une astuce de la raison pour ne pas avoir à admettre sa propre impuissance : prise au sérieux, l'origine est *sans raison* – et face à elle, la raison est réduite au silence. Le problème n'est pas que nous ignorons la raison de ce qui est. C'est qu'il ne peut pas y en avoir, car s'il y en avait une, la raison exigerait la raison de la raison de ce qui est. En admettant que tout soit rigoureusement déterminé, que nous soyons les rouages d'une gigantesque mécanique, comment rendre compte de l'existence de cette machine ? Nous voulons bien admettre que tout soit descriptible de manière mécanique, une telle description manque le fait qu'*il y ait* cette machine. Aucun système ne peut rendre compte de ce « il y a ». Il n'y a pas de modèle qui rende compte du fait qu'il y ait quelque chose plutôt que rien. Autrement dit affirmer qu'il n'y a aucun surgissement revient à manquer le *fait de l'existence.*

813 Ici nous faisons abstraction d'éventuelles différences de sens entre les termes de « cause », « raison », et « condition ». Nous les considérons comme des synonymes.

3.6.2. Quelle place pour des événements sans cause ?

Une fois que la pensée est contrainte d'admettre qu'il existe, *ne serait-ce qu'une seule chose* sans raison – comme l'origine de ce qui existe, autrement dit le fait qu'il y ait quelque chose plutôt que rien –, alors rien n'empêche de penser que *d'autres choses* soient dépourvues de raison – comme l'apparition de la vie ou de la conscience. Une seule exception au principe de raison l'amène à perdre le statut de principe et laisse planer la perspective d'autant d'*événements sans cause* dont il faudrait accepter qu'ils ont eu lieu sans que la science ne puisse en dire autre chose qu'« *ignorabimus* ». Mais encore faut-il admettre ce premier *ignorabimus.* Les philosophes et les poètes ont beau répéter que « la rose est sans pourquoi[814] », les chercheurs ne semblent pas en mesure de le recevoir, puisque l'argument ne se formule pas dans le seul langage qu'ils veulent bien entendre, celui des preuves formelles. La considération du « il y a » est hétérogène aux formalisation logiques ou informatiques. Au mieux, ils se comportent comme s'il n'y avait qu'*une seule origine,* un hypothétique moment d'avant le *big bang,* le reste découlant ensuite selon une causalité rigoureuse. Mais s'il y a une exception avec l'émergence de la matière, pourquoi n'y en aurait-il pas deux autres, le surgissement de la vie et de la conscience ? Quelle place, alors, donner à ces événements sans cause ? L'apparition de la matière, de la vie et de la conscience seraient-elles trois exceptions à un ordre du monde par ailleurs rigoureusement nécessaire ? Est-il possible de *compartimenter* les événements contingents de manière à ce que tous les autres événements restent sous l'autorité rassurante du principe de raison ? Comment tenir une vision du monde où il y aurait *alternance* entre contingence et nécessité ? Admettre quelques exceptions à la nécessité revient à peindre un monde où il y aurait des miracles – sans qu'ils soient pour autant attribués à un Dieu. Dès lors qu'un, deux, ou trois événements sont *sans raison,* rien n'assure que *chaque événement* ne participe pas de la même contingence. Autrement dit, l'apparition de la matière, de la vie et de la conscience pourraient n'être que les exemples les plus spectaculaires d'un hasard qui qualifie tout ce qui vient à l'existence – naissances, rencontres, événements. Cela rend le monde plus simple et plus difficile à penser, mais permet de mieux sentir ce que nous entendons lorsque nous parlons de créativité ou de liberté – plutôt que de postuler une originalité sans origine, comme le font, dans la lignée de l'article de 1950, les chercheurs en intelligence artificielle.

814 Martin Heidegger, *Le principe de raison, op. cit.*, p. 102-109.

Cela permet d'éviter les apories que produisent un paradigme de l'intervention (voir chapitre 3.2. Donner du jeu : le hasard). Il n'y a pas d'« interventions » inexplicables de la vie, et de la conscience « dans » une nécessité préexistante. Il y a, de bout en bout, intervention inexplicable – hasardeuse – de l'existant. Le surgissement *ex-nihilo* est permanent. Prise dans tous ses détails, sans rien négliger, chaque situation est une conjonction, un collage inédit. Elle ne s'est jamais produite avant et ne se reproduira jamais. Elle est neuve, singulière, sans modèle, *originale* (voir section 2.6.2. L'intuition de l'instant). Autrement dit, un seul événement singulier et hasardeux, comme le fait qu'il existe quelque chose plutôt que rien, ouvre la voie à une manière de voir où tout ce qui existe est également singulier et hasardeux.

3.6.3. Contingence des phénomènes de l'esprit

Nous inférons des lois à partir des régularités que nous observons, nous nous donnons des règles collectives, nous contractons des habitudes individuelles, mais tous ces efforts vers la stabilité ne peuvent effacer que notre expérience la plus intime, celle de l'intuition et plus généralement de nos *états d'esprits,* est contingente. La raison a beau faire *comme si* aucune nouveauté n'intervenait, elle ne cesse de se heurter à des phénomènes sans cause apparente, à commencer par sa propre activité. C'est toute l'ironie du projet d'intelligence artificielle, qui revient à chausser les lunettes du mécanisme pour scruter les phénomènes qui le contredisent le plus : l'esprit humain et sa liberté, ou tout du moins sa « créativité ». L'idée apparemment folle qu'il y a des événements sans cause est la chose la plus banale, pour peu que, comme Lewis Carroll, on prenne le temps d'observer et de décrire nos états mentaux. Si une partie de ceux-ci répond au principe de raison (« je suis de bonne humeur *parce que* j'ai bien dormi »), chacun aura fait l'expérience de ces moments où, alors que toutes les conditions sont rassemblées pour susciter un état mental, celui-ci ne vient pas : insomnie, panne d'inspiration, impuissance sexuelle, indifférence devant une œuvre, déception face à son plat préféré… Pire encore, tous les efforts que nous pourrions faire pour les causer ne font que les éloigner d'autant. Romain Graziani nomme « états réfractaires à la conscience réflexive » ces états mentaux que l'on empêche d'advenir à trop les vouloir, ainsi que ceux que l'on attire à force de les éviter. Il critique la

célébration sans réserve de la volonté (« l'éthique musculaire ») et décrit les « crampes » qui adviennent lorsque l'humain se traite comme une machine[815].

Des remarques analogues s'appliquent à la joie. Il y a ces journées où, bien que toutes les conditions soient au rendez-vous (il fait beau, c'est les vacances, les amis sont là), la bonne humeur ne vient pas. Rien ne semble pouvoir déloger une morosité dont la présence n'a pas de raison identifiable. Les causes habituellement agissantes n'agissent plus. Pire, il arrive qu'elles s'inversent : cela me contrarie encore plus d'être contrarié *alors qu'il fait si beau.* Tout ce qui pourrait causer une meilleure humeur aggrave la mauvaise en soulignant à quel point la bonne humeur fait défaut : la musique m'agace, les gentillesses me hérissent… Jusqu'à ce qu'un « rien » – un voisin marche dans une crotte, ou un oiseau se pose sur la table – vienne dissiper la morosité qui m'affectait et laisse la place à une humeur plus clémente. Clément Rosset n'aura eu de cesse de montrer que *la joie est sans raison.* À bien y réfléchir – que l'on pense au vieillissement, à la mort, à l'extinction de l'espèce humaine ou à celle du soleil – il n'y a *aucune raison* d'être joyeux. Pourtant la joie est là, comme peut le rappeler un simple morceau de musique. Et cette joie est d'autant plus forte, plus « profonde », si elle est sans raison, si c'est une joie « folle[816] ».

La raison est aussi impuissante à causer mécaniquement la joie qu'à se débarrasser des phobies, comme se plaît à le rappeler Pascal en imaginant qu'un champion de la raison (« le plus grand philosophe du monde »), s'il était au-dessus d'un précipice « sur une planche plus large qu'il n'en faut », c'est-à-dire hors de danger, n'en aurait pas moins le vertige[817]. Le philosophe a beau savoir qu'il ne risque rien, cela ne l'empêche pas de ressentir la peur du vide. Tout se passe « comme si son vertige était hors de portée de sa philosophie[818] ». Toutes les idées qu'il pourrait avoir sur la situation sont dépourvues de pouvoir causal. Et s'il semble bien y avoir une cause au vertige – le vide ou plutôt l'imagination du vide – celle-ci n'est pas pour autant systématique. Il arrivera que le vertige ne se manifeste pas là où on l'aurait attendu – et notamment les jours où l'on pourrait souhaiter qu'il se manifeste, à l'instar d'un symptôme qui disparaît le jour où on se décide à le montrer à un médecin.

---

815 Romain Graziani, *L'usage du vide, Essai sur l'intelligence de l'action, de l'Europe à la Chine,* Paris, Gallimard, 2019.

816 C'est un thème récurrent dans l'œuvre de Clément Rosset et l'argument principal de *La force majeure*, Paris, Éditions de Minuit, 1983. La plupart des philosophes (du souverain bien d'Aristote à la joie de Spinoza) ayant montré le contraire, que la philosophie et donc la raison pouvaient apporter la joie, cela a valu à Clément Rosset le qualificatif d'« antiphilosophe ».

817 « Le plus grand philosophe du monde, sur une planche plus large qu'il ne faut, s'il y a au-dessous un précipice, quoique sa raison le convainque de sa sûreté, son imagination prévaudra. Plusieurs n'en sauraient soutenir la pensée sans pâlir et suer. » Pascal, *Pensées*, Paris, Bordas, 1966, pensée 82 (Brunschvicg), p. 54.

818 L'expression est de Pierre Cassou-Noguès, qui a publié un travail en ligne sur les phobies, voir substancejournal.lmu.build/phobic-postcards/index

Avec l'effet placebo, la médecine se trouve devant un effet dont on ignore la cause : « Des effets physiques, qui relèvent donc de la biologie, interprétée par la physico-chimie, apparaissent sans cause du même ordre[819]. » Pourtant, « le placebo non scientifique fonde la scientificité de tout médicament » puisque « pour vérifier qu'une sustance est réellement active dans telle affection, il est nécessaire de comparer ses effets à ceux d'une substance neutre[820] », un médicament doit donc être comparé avec l'effet placebo afin de se voir attribuer un pouvoir causal. Il serait peut-être plus pertinent de parler de *bruit* placebo et de renoncer à l'interpréter, mais c'est le terme *d'effet* qui a prévalu, invitant à poursuivre la quête d'une cause expliquant l'effet. Les candidats à la cause se multiplient (le rituel médical, les croyances des patients…) sans qu'aucune ne parvienne à tenir pleinement son rôle. On a beau essayer de supprimer la cause suspectée, en enlevant aux médecins leurs blouses blanches ou bien en informant le patient qu'il s'agit d'un placebo, son ampleur diminue mais l'« effet » placebo persiste.

Pour les scientifiques, le plus irritant n'est pas tant de ne pas trouver la cause, puisqu'ils ne doutent pas qu'ils finiront par avoir « gain de cause », que d'observer la débauche d'hypothèses saugrenues que cela suscite dans le public, voire chez leurs confrères, et l'adhésion irréfléchie qu'elles rencontrent : phénomènes quantiques, mémoire de l'eau, pouvoirs psychiques du médecin ou du patient, arbitraire divin, anges ou démons… Là où l'explication scientifique est défaillante, les candidats se bousculent pour proposer une explication *plus mauvaise encore*. Le phénomène inexpliqué donne libre cours à l'imagination la plus folle, comme *s'il fallait absolument une explication* – une hypothèse farfelue ou ridicule valant apparemment mieux que le silence. La rigueur voudrait pourtant que l'on s'abstienne, qu'on s'en tienne à admettre qu'il n'y a pas d'explication – quitte à débattre ensuite sur le sens à donner à cette absence, qu'on veuille croire que l'explication est à venir – il n'y a *pas encore* d'explication – ou qu'elle pourrait tout aussi bien faire défaut – il n'y a *pas du tout* d'explication.

Ce défaut d'humilité donne lieu à un malentendu persistant. Lorsque la contingence des phénomènes psychiques intervient, on dira que « c'est dans la tête » du patient, comme s'il y avait d'un côté le corps, aussi prévisible et déterminé que le cosmos selon la physique du dix-septième siècle, et de l'autre l'esprit, ou *a minima* l'imagination, une folle du logis venant perturber les opérations. Il s'agit là, à nouveau, du paradigme de l'« intervention » du contingent dans un cours par ailleurs nécessaire : l'esprit, logé dans la tête, interviendrait parfois

819 François Roustang, *La fin de la plainte,* in *Jamais contre d'abord, la présence d'un corps,* Paris, Odile Jacob, 2019, p. 245. Le chapitre 12 est consacré à « L'effet placebo, conséquence d'un rite », p. 245-270.
820 *Ibid,* p. 247.

dans les affaires du corps, amenant à des symptômes sans cause ou à une rémission – comme Dieu envoyant des miracles ou des punitions. Penser les choses de cette manière, c'est se condamner à ne rien comprendre du corps, ni de l'esprit. Les psychiatres ont beau s'acharner à inscrire les phénomènes qu'ils traitent dans la matérialité de processus corporels comme la chimie des neurotransmetteurs, cela ne les rendra pas forcément plus prévisibles pour autant. Le corps naît du même hasard que l'esprit. Il faut inverser la perspective habituelle : il n'y a pas « intervention » du hasard (souvent rapporté à l'esprit) dans des phénomènes par ailleurs déterminés (généralement rapportés à la matérialité du corps), mais il y a formation de stabilités transitoires depuis un fond hasardeux. Les corps qui nous composent, ou qui composent le cosmos, se sont accrochés d'une manière qui aurait pu être différente. En se rencontrant, ils ont *contracté des habitudes* plus ou moins durables. Par exemple, le système solaire peut être conçu comme une belle machine parfois perturbée par la rencontre de corps extérieurs, ou bien comme la rencontre de corps variés ayant donné lieu aux mouvements réguliers que nous connaissons. La régularité de ces mouvements constitue une structure, mais celle-ci reste éphémère. Nous pouvons connaître ces structures et les utiliser à notre profit, mais sans savoir jusqu'où elles tiendront. Concrètement, cela se traduit par une pratique qui renonce à la certitude. Pour la plupart des médecins, c'est la chose la plus évidente du monde, et il ne sert à rien de le dire. Que dire d'ailleurs, pour parler de l'ignorance ? Tout l'espace du discours est occupé par quelques autres qu'une illusion de certitude – permise par un éloignement de la pratique – rend bavards. Ils font miroiter la perspective d'une médecine débarrassée de l'arbitraire, devenue « scientifique » de bout en bout, et l'assortissent de promesses plus ou moins explicites sur l'immortalité, la résurrection des morts et le téléchargement des esprits. Si la médecine, en se définissant comme un « art », a toujours eu la sagesse de se garder des illusions de certitude véhiculées par une certaine idée de la science, c'est contrainte par l'expérience de la contingence des corps. Il aura fallu un long chemin aux autres disciplines, à commencer par la physique, qui a incarné pour les autres le modèle de la certitude scientifique, pour réaliser qu'il lui fallait également faire une place au hasard, montrant que ce n'est pas seulement la médecine, mais toute la science, qui doit faire une place à l'idée d'« art », que l'arbitraire n'est pas l'apanage de l'humain.

3.6.4. La science et le temps

Pour Platon, comme pour Aristote, l'horizon de la connaissance est hors du temps – ce sont les Idées, ou le premier moteur. La recherche philosophique revient à délaisser le « domaine du devenir, du multiple, de l'instable de l'illimité, de l'opinion biaisée et flottante » au profit du « domaine de l'être, de l'un, de l'immuable, du limité, du savoir droit et fixe[821] », autrement dit à se tourner vers *ce qui demeure* plutôt que *ce qui devient.* Dans le monde décrit par Aristote, les mutations sont bornées « par le haut et par le bas » : le monde supralunaire[822] (« en haut ») ainsi que l'*hupokaimenon*[823] (« en bas ») sont permanents et incorruptibles. En définissant les « accidents » comme ce qui advient à une substance elle-même nécessaire et en les cantonnant au monde sublunaire, Aristote influence deux mille ans d'une vision du monde où la variabilité des vies terrestres est l'exception dans un cosmos régi par la nécessité.

Cette conception du cosmos est mise à mal par le geste de Galilée. En tournant sa lunette vers le ciel, il accumule les observations – tâches du Soleil, bizzarerie de la Lune – invalidant l'incorruptibilité des astres. Lorsque Tycho Brahé établit que les comètes sont des objets supralunaires, il devient indéniable qu'Aristote avait tort : les cieux ne sont pas éternels. Si le ciel galiléen est réinscrit dans le temps, il n'est pas pour autant livré à la contingence. Le monde supralunaire – comme le monde sublunaire – est soumis aux changements, mais dans les deux mondes, ces changements suivent des lois éternelles. Tout comme la trajectoire des comètes, l'univers « est écrit en langue mathématique[824] » et l'ensemble des mouvements célestes est prévisible. L'immuabilité du cosmos est ainsi remplacée par l'immuabilité des lois qui régissent ses mouvements. L'image de la machine, qui hante la science et la philosophie du dix-septième siècle, permet de réconcilier le mouvement et la permanence : une machine passe par différents états mais son plan ne change pas. Les comètes et autres événements cosmiques suivent le plan nécessaire de la mécanique céleste, un plan que les physiciens s'attachent à deviner par le truchement de l'observation et des mathématiques. Ainsi, le bouleversement galiléen détrône la cosmologie aristotélicienne mais réaffirme la permanence du monde. L'objet de la recherche

821 Marcel Detienne et Jean-Pierre Vernant, *Les ruses de l'intelligence, la* mètis *des Grecs,* Paris, Flammarion, 1974, p. 11.
822 Dans le monde selon Aristote, ce qui est sous la Lune (monde « sublunaire ») est soumis au changement, alors que ce qui est au-dessus (monde supralunaire) ne l'est pas.
823 Le sujet, ou le substrat du changement.
824 Galilée, Christiane Chauviré, *L'Essayeur de Galilée*, Paris, Les Belles Lettres, 1979, p. 141.

reste ce qui demeure, le plan fixe qui régit le changement. L'horizon de la connaissance est encore la contemplation de formes éternelles.

Après avoir porté des fruits considérables, le paradigme mécaniste s'étiole peu à peu et les physiciens renoncent au paradigme d'une nature prévisible. En 1927, si Einstein juge bon d'insister sur le fait que « Dieu ne joue pas au dés[825] », c'est que les progrès de la physique semblent montrer le contraire, aussi bien dans l'infiniment grand que dans l'infiniment petit. Avec le problème des trois corps, il est démontré que les mouvements d'un système aussi « simple » que notre système solaire sont imprévisibles[826]. Les atomes, substrats matériels des accidents, réputés indivisibles et incorruptible, se trouvent avoir des composants, dont l'un – le neutron – ne dure que quinze minutes alors que l'autre – le proton – n'est considéré comme éternel que parce-que sa durée de vie est supérieure à l'âge de l'univers. Ces composants sont eux-mêmes composés de particules élémentaires que les physiciens ont pris l'habitude de voir apparaître et disparaître au sein de leurs accélérateurs[827]. Enfin certaines constantes fondamentales sont suspectées d'avoir varié par le passé[828] – ce ne sont pas des *constantes*. Au fil des découvertes, l'univers semble moins dépendre de lois nécessaires que d'une histoire contingente.

Au-delà du clivage entre physique relativiste et physique quantique, qui met à mal la discipline en tant que modèle de « science dure », ces découvertes la contraignent à tourner son regard vers le devenir et à remettre en question le statut des régularités qu'elle a trop rapidement qualifiées de « lois ». De même que lorsqu'on refroidit les atomes d'un aimant, ceux-ci prennent spontanément une direction commune, il est possible qu'à mesure que l'univers se refroidissait la matière ait pris certaines « orientations ». Autrement dit, si les particules semblent suivre des « lois » communes, c'est parce qu'elles ont la même histoire. Les mêmes événements contingents, ou « brisures de symétrie », leur ont fait prendre certaines « directions » communes qui apparaissent aujourd'hui comme nécessaires – la dépendance au chemin (*path dependency*) les empêchant de changer. Mais celles-ci ayant pu être autre, elles n'ont pas la nécessité qu'on leur prête. Rien n'empêche de penser que, tout comme il est possible de démagnétiser un aimant, certaines constantes perdraient leur efficace si l'univers

825 Lors du congrès de Solvay. Manjit Kumar, *Le Grand Roman de la physique quantique, Einstein, Bohr… et le débat sur la nature de la réalité*, Paris, Champs Flammarion, 2012, p. 389.

826 Nous empruntons cet exemple à Guiseppe Longo, « Cercles vicieux, mathématiques et formalisations logiques » *Mathématiques et Sciences Humaines*, n°151, 2000.

827 Nous empruntons ces exemples à la conférence de Pierre Lena au Collège de France, « D'un univers statique à un univers en devenir », 25 janvier 2007, https://www.college-de-france.fr/site/anne-fagot-largeault/course-2007-01-25-10h30.htm page consultée le 28 mai 2020.

828 John D. Barrow, « Varying Constants », *Philosophical Transactions of the Royal Society*, 363, p. 2139-2153.

venait à entrer en contraction et prendraient d'autres valeurs, avec d'autres effets, si l'univers entrait à nouveau en expansion. En d'autres termes, les « lois » sont le produit d'une inertie dont rien ne nous dit qu'elle n'est pas réversible.

On doit à Pierce[829] une réflexion sur la possibilité d'inscrire l'ensemble des « lois de la nature » dans le temps, avec l'idée que ces lois de la physique pourraient être « le résultat de l'évolution », tout autant que les lois de l'esprit et de la vie. Les lois de la physique seraient aussi « plastiques » que les lois du monde organique ou celles de la vie de l'esprit, mais à des échelles de temps différentes. Pour des phénomènes comme la gravitation l'échelle est si grande « qu'on ne peut rien y trouver, pas même un semblant d'irrégularité. » Dès lors il convient d'étudier les lois dans les domaines où elle est visible pour nous, c'est-à-dire l'esprit humain ou le monde organique : « L'esprit humain est la plus plastique de toutes les choses ; ensuite vient le monde organique, le monde du protoplasme. » Pour Pierce, si ces lois évoluent, elles le font également selon une loi. « Si les lois de la nature sont le résultat de l'évolution, cette évolution doit procéder selon un certain principe ; et ce principe sera lui-même de la nature d'une loi. » Or pour que cette loi, comme toute les autres, soit inscrite dans le temps, il faut qu'elle soit « telle qu'elle puisse évoluer ou se développer. » Cette loi des lois est dans le fait de *former* des lois, que Pierce appelle la « tendance généralisante ». C'est « la grande loi de l'esprit, la loi de l'association, la loi de la prise d'habitude. » Et Pierce de conclure : « J'en suis donc venu à l'hypothèse que les lois de l'univers se sont formées sous l'effet de la tendance universelle de toute chose à se généraliser et à prendre des habitudes[830]. »

829 Charles Sanders Pierce, *Le raisonnement et la logique des choses, Les conférences de Cambrigdge (1898),* Paris, Le Cerf, 1995, p. 309-310. Le texte est cité et commenté par Anne Fagot-Largeault, *Ontologie du devenir, L'évolution, l'univers et le temps,* Paris, Odile Jacob, 2021, p. 270-273.

830 « Si les lois de la nature sont le résultat de l'évolution, cette évolution doit procéder selon un certain principe ; et ce principe sera lui-même de la nature d'une loi. Or il doit s'agir d'une loi telle qu'elle puisse évoluer ou se développer. [...] Où irons-nous la chercher ? Nous ne pouvons pas nous attendre à la trouver dans des phénomènes comme la gravitation, où l'évolution est si proche de la limite définitive qu'on ne peut rien y trouver, pas même un semblant d'irrégularité. Il nous faut plutôt chercher cette tendance généralisante dans des domaines de la nature où sont encore à l'œuvre plasticité et évolution. L'esprit humain est la plus plastique de toutes les choses ; ensuite vient le monde organique, le monde du protoplasme. Or la tendance généralisante est la grande loi de l'esprit, la loi de l'association, la loi de la prise d'habitudes. Dans tout protoplasme actif, nous trouvons aussi une tendance à prendre des habitudes. J'en suis donc venu à l'hypothèse que les lois de l'univers se sont formées sous l'effet de la tendance universelle de toute chose à se généraliser et à prendre des habitudes ». Charles Sanders Pierce, *Le raisonnement et la logique des choses, Les conférences de Cambrigdge (1898),* Paris, Le Cerf, 1995, p. 309-310.

3.6.5. Le fantôme du mécanisme physique

Déchirée entre les paradigmes relativistes et quantiques, contrainte de faire une place à l'histoire, la physique a pourtant été maintenue comme archétype de la « science dure » consistant à identifier par l'observation des lois dont la meilleure expression serait mathématique. Au lieu de tourner leur regard vers le devenir, les autres disciplines (biologie, psychologie, sociologie, économie…) perpétuent un tropisme vers les « modèles » hors du temps et grandissent à l'ombre d'une idée de la physique que la physique a progressivement cessé d'incarner. Ainsi au dix-neuvième siècle, la sociologie naissante se veut une « physique sociale », selon les termes d'Auguste Comte. Les statistiques donnent à voir « l'effrayante exactitude avec laquelle les crimes se reproduisent[831] » et laissent penser que la société est un corps artificiel organisé selon des lois que la sociologie commence tout juste à découvrir. De même, au vingtième siècle, l'« organisation scientifique du travail » aborde l'activité sous le prisme de la physique, comme des échanges de masses et d'énergie. Le travail est vu comme un ensemble de mouvements réglés dont le modèle est l'horloge. « L'ouvrier des *Temps Modernes* se trouvera réduit à un jeu de forces physiques asservi à la cadence de la chaîne de production. Son corps devra se plier au modèle de l'horloge cher à Hobbes et à La Mettrie[832]. »

À partir des travaux de Mendel, le même mécanisme s'empare de la biologie et l'amène à répéter le geste de la physique classique : l'évolution récemment découverte est assignée à un cadre immuable, un ensemble de lois qui ne changent pas et qui permettent à la biologie de se prendre pour une science exacte. « Au milieu du vingtième siècle, il semblait qu'en biologie les lois avaient triomphé de la perspective historique[833] ». Au cours de la deuxième moitié du vingtième siècle, alors que l'interprétation déterministe de la génétique gagne le grand public, une succession de découvertes vient l'invalider. Il est montré par exemple que le même gène peut fabriquer des protéines différentes, ou que des fragments d'ADN peuvent se déplacer (« gènes sauteurs »). Enfin, les expériences de clonage mettent en lumière l'importance des facteurs non génétiques. Au tournant du siècle, Evelyn Fox Keller récapitule cette épopée dans

831 Olivier Rey, *Quand le monde s'est fait nombre,* Paris, Stock, 2016.

832 Alain Supiot, *La Gouvernance par les nombres, Cours au Collège de France (2012-2014),* Paris, Fayard, 2015, p. 41.

833 Anne Fagot-Largeault, *Ontologie du devenir,* cours au Collège de France du 14 février 2008, seconde partie, minute 25, https://www.college-de-france.fr/site/anne-fagot-largeault/course-2007-2008.htm, page consultée le 20 mai 2020.

un ouvrage qui signe l'acte de décès du concept de gène[834]. Notre évolution se sert du gène comme on se sert d'un livre dans une bibliothèque. C'est une illusion de penser que le gène est un « programme » qui maîtrise notre développement. Cependant, la mauvaise réception de l'ouvrage par une partie des scientifiques montre la persistance de l'erreur qu'elle dénonce. Les biologistes ont été pris au piège du rêve du « gène maître ». L'évolution est systémique au sens de Bateson, c'est-à-dire que les composants du système s'entre-déterminent sans que l'un ne « commande » les autres. Il est préférable de penser qu'« aucune partie de ce système intérieurement (inter) actif ne peut exercer un contrôle unilatéral sur le reste ou sur toute autre partie du système[835]. »

Hubert Dreyfus note que la même fascination pour les succès de la physique motive l'optimisme de la première génération de chercheurs en intelligence artificielle[836]. Comme le reste de l'univers, les mouvements du cerveau s'écrivent en langue mathématique. Rien ne s'oppose à ce que soit découvert le « système » qui régit les mouvements entre ses différentes parties comme a été découvert le « système » qui régit les mouvements des planètes. Reste à savoir sur quelle partie élémentaire de ce « système » il faut mettre l'accent : les neurones ou bien l'information.

À l'instar d'Hubert Dreyfus pour l'intelligence artificielle ou d'Evelyn Fox Keller pour la génétique, de nombreux auteurs[837] ont démontré que le paradigme de la physique classique mécaniste n'est pas plus pertinent pour leur objet qu'il ne l'a été pour le cosmos. Qu'ils aient dû se donner la peine de le montrer prouve cependant que malgré les évolutions des différentes disciplines, la physique classique reste l'idéal des sciences de la nature. La plupart des disciplines scientifiques continue d'être hantée par le fantôme de la physique classique.

3.6.6. Alphabêtise de la science

D'après Giuseppe Longo et Pierre-Emmanuel Tendero, si le modèle mécaniste a une telle force de fascination, c'est parce qu'il répète l'émerveillement suscité par l'invention de l'alphabet, et avec lui, l'illusion de pouvoir tout exprimer grâce à un ensemble fini de composants

834 Evelyn Fox Keller, *The Century of the Gene*, Cambridge MA, Harvard University Press, 2000, cité par Anne Fagot-Largeault, *op. cit.*
835 Gregory Bateson, *op. cit.*, p. 272.
836 Hubert Dreyfus, *Intelligence artificielle : mythes et limites*, *op. cit.*, p. 239-247.
837 Par exemple, les travaux de Giuseppe Longo sur la biologie ou ceux d'Isabelle Stengers sur la chimie.

élémentaires et de règles de composition[838] – illusion pourtant démentie par l'existence des idéogrammes. Avec la physique classique, le paradigme alphabétique laisse croire qu'il est possible d'exprimer l'ensemble des phénomènes de l'univers dans un système d'écriture fini. Les auteurs remarquent que le même rêve se retrouve en mathématiques, lorsque le projet de Hilbert entreprend l'élaboration d'un seul système formel capable de démontrer ou invalider toutes les propositions des mathématiques, ou encore en biologie, lorsque la découverte de l'ADN est interprétée comme « l'alphabet du vivant ».

Répandue parmi la première génération de chercheurs en intelligence artificielle, Hubert Dreyfus appelle cette croyance le « postulat ontologique » et lui signale de prestigieux précurseurs, à commencer par Platon, qui exige que « tout savoir soit exprimé sous formes de règles ou de définitions » à partir de l'hypothèse que « tout peut se réduire à des éléments simples, auxquels des règles s'appliquent[839] ». C'est Leibniz qui en affiche le plus clairement l'ambition, puisqu'il évoque « un alphabet des pensées humaines » et rêve à ses indéniables avantages, parmi lesquels celui de rapporter les querelles d'opinion à des questions de manipulation d'objet – et donc d'y mettre fin. Dans la même lignée, Hubert Dreyfus signale également les éléments discrets de Hume, l'atomisme logique de Russell et surtout le *Tractatus* de Wittgenstein qui « définit le monde comme un système d'atomes de faits qu'il est possible d'exprimer sous forme de propositions logiquement indépendantes[840] ».

Le paradigme alphabétique a été critiqué par de nombreux philosophes, à commencer par Heidegger attaquant la « pensée calcul » ou Merleau-Ponty critiquant « le préjugé du monde ». Là encore, Wittgenstein se distingue en devenant le pourfendeur impitoyable d'une idée qu'il avait d'abord défendu avec ferveur. L'histoire de la philosophie verrait d'un côté les « philosophes de système[841] » qui, à la suite des pythagoriciens, puis des tenants d'une *clavis universalis*[842], postulent que le monde a une structure combinatoire cachée, et de l'autre ceux que Clément Rosset appelle les philosophes tragiques[843], pour qui la structure postulée est

838 Giuseppe Longo, Pierre-Emmanuel Tendero. « L'alphabet, la Machine et l'ADN : l'incomplétude causale de la théorie de la programmation en biologie moléculaire ». in Paul-Antoine Miquel, *Biologie du XXIe siècle : évolution des concepts fondateurs*, Louvain la Neuve, DeBoeck, 2008.

839 Hubert Dreyfus, *op. cit.*, p. 269.

840 *Ibid.*

841 Jacques Bouveresse, *Qu'est-ce qu'un système philosophique ?, Cours 2007 et 2008,* Paris, Collège de France, 2013.

842 Ainsi que toute la tradition de l'art de mémoire : « l'Art n'est pas une technique liée aux finalités du discours rhétorique, mais principalement, l'instrument à utiliser pour fonder un édifice dont les structures constituent le reflet des structures de la réalité. Les règles de la mémoire, ainsi que les techniques combinatoires, trouvent leur justification dans le postulat, clairement admis, d'une totale correspondance entre les symboles et les res, entre les ombres et les idées, entre les sceaux et les raisons qui président aux articulations du monde réel. » Paolo Rossi, *Clavis Universalis,* Grenoble, Jérôme Millon, 1993, p. 105.

843 Clément Rosset, *Logique du pire,* Paris, Presses Universitaires de France, 1971.

comme la physique contemporaine : la cohérence de l'ensemble, tout autant que la simplicité des atomes censés le constituer, ne résistent pas à l'examen. Clement Rosset mentionne ainsi Pascal, Montaigne, ou Epicure. C'est le statut accordé au langage qui les distingue : pour les premiers, le langage peut représenter les phénomènes car il partage la même structure qu'eux – il y a un alphabet du réel ; tandis que pour les seconds, la singularité du réel rend vaine toute tentative de représentation.

3.6.7. Singularité de la réalité et impuissance de la représentation

C'est la singularité des objets qui nous rend impuissants à les représenter. Clément Rosset a consacré un petit ouvrage à cette notion de singularité, qui est pour lui indispensable à toute évocation du réel[844]. Toute représentation, ou « réplique », d'un objet, se construit à partir de notions communes et par comparaison avec d'autres objets connus. Nous ne connaissons un objet que par différence avec les autres objets. Par conséquent, ce qui fait le propre d'un objet – ce en quoi il *n'y a rien de tel* – est impossible à représenter : aucune comparaison n'est possible puisqu'il n'existe aucun comparant au moyen duquel se le figurer. La singularité ne peut être médiatisée : « la différence se perçoit mais non ce dont elle diffère, c'est-à-dire l'identité[845] ». C'est « le caractère impensable et indescriptible du *même* dès lors qu'il n'est aucun *autre* pour en rendre raison[846] ; » Cette invisibilité des objets singuliers n'est pas l'apanage d'une catégorie particulière d'objets. C'est ainsi qu'est l*e réel* en tant qu'« ensemble non clos d'objets non identifiables[847] » :

> L'objet réel est en effet invisible, ou plus exactement inconnaissable et inappréciable, précisément dans la mesure où il est *singulier,* c'est-à-dire tel qu'aucune représentation ne peut en suggérer de connaissance ou d'appréciation par le biais de la *réplique*[848].

Clément Rosset le reformule à plusieurs reprises : « Le réel est ce qui est sans double, soit une singularité inappréciable et invisible parce que sans miroir à sa mesure ». Il va jusqu'à affirmer que « l'invisibilité du réel est un caractère constitutif du réel. Et cette invisibilité procède de la

844 Clément Rosset, *L'objet singulier,* Paris, Les Éditions de Minuit, 1979.
845 *Ibid*, p. 20-21.
846 *Ibid*, p. 16.
847 *Ibid*, p. 22.
848 *Ibid*, p. 15.

singularité de l'objet réel. En tant que singulier, il est hors de toute comparaison, de toute mesure, de toute raison[849]. » Hors de toute *ratio* donc de tout équation, le réel est singulier comme on dit en mathématiques ou en physique qu'un point est singulier. N'importe quel objet existant, (« réel » dirait Clément Rosset), même le plus petit, ne peut être représenté dans l'ensemble de ses caractéristiques, car il a des caractéristiques singulières, celles qu'il est seul à porter, ou parce qu'il est défini par l'ensemble de ses relations aux autres objets : il faudrait restituer l'ensemble du monde pour le représenter « fidèlement ».

Il ne s'agit pas de postuler qu'une chose singulière présente *une ou plusieurs parties* irreprésentables. Cela nous amènerait à des paradoxes similaires au « paradoxe du contre-argument » présenté en section 1.1.10 au sujet de la conjecture de Dartmouth. S'il s'agit d'une caractéristique, elle devrait pouvoir être définie et représentable. S'il est possible de l'isoler, il est possible de la décrire et d'en faire le support d'une analogie avec autre chose : ce rouge, cette courbe, se retrouvent dans tel ou tel autre objet. Seule la chose prise sous l'aspect de *la totalité de ses caractéristiques*, n'a rien de tel. C'est l'ensemble de ses parties, le *collage* qui la constitue, qui est singulier, à commencer par le collage avec un lieu et une époque. Pour décrire ce qui constitue un objet singulier – et donc *ce* qui fait sa singularité – il ne faut rien négliger. La singularité se constitue dans la rencontre inédite entre une infinité de *détails*. Pour restituer une chose singulière, il faut donc considérer tout ce qui la constitue, et comment tout ce qui la constitue a été constitué – tout ce qu'elle rencontre et tout ce qu'a rencontré ce qu'elle rencontre. La singularité d'une chose, c'est la totalité de ce qui existe pris depuis le point de vue de cette chose. Reproduire une chose singulière, c'est reproduire l'ensemble de ce qui a existé et conspiré à l'existence de cette chose, c'est recommencer tout l'univers.

Par contigüité temporelle et spatiale, toutes les entités du monde sont *parties* prenantes de l'affaire et fournissent un ingrédient plus ou moins déterminant du collage singulier. Pour ne rien omettre, c'est la totalité de ce qui existe qu'il faudrait considérer. L'impossibilité à reproduire un objet singulier ne tient donc pas à l'irreprésentabilité d'*une ou plusieurs* caractéristiques : il n'y a pas d'aspect indescriptible d'une chose dont on pourrait dire « voilà la caractéristique que vous n'arriverez pas à décrire, voilà ce qui la rend singulière » – tout comme, ainsi que cela a été formulé à l'occasion du séminaire de Dartmouth, il n'y a pas de caractéristique de l'intelligence qu'il serait impossible de décrire sous la forme d'une machine. Mais chacune des caractéristiques peut être descriptibles sans que l'ensemble le soit. Pour le résumer en une phrase : une chose singulière peut être irreprésentable (et elle l'est), sans

849 *Ibid*, p. 16.

qu'aucune de ses caractéristiques ne soit, elle, irreprésentable. La chose singulière ne peut être reproduite car ses parties incluent l'infinité des parties du monde. Quand bien même ces parties seraient finies, nous n'aurions pas le temps de les reprendre toutes. Quand bien même nous aurions le temps, nous ne pourrions reproduire la conjonction particulière offerte par un instant donné (voir section 2.6.2 L'intuition de l'instant). Aussi, la seule représentation *complète* d'une chose, incluant toutes les contiguité avec les autres choses, et surtout son lieu et son époque, c'est la chose elle-même.

3.6.8. Le deuil du miroir

Nés du désir de représenter une hypothétique intelligence universelle, les outils inventés par les chercheurs en IA ont progressivement évolués vers la singularité des objets auxquels ils sont appliqués. Autrement dit, l'histoire de l'intelligence artificielle évolue, comme celle de la physique, suivie par la chimie, la biologie, la génétique, vers une prise en compte de la singularité des objets décrits et un deuil de l'idéal du modèle hors du temps. Mais le fantôme du mécanisme ne semble pas pour autant s'y être dissipé. *Dans la pratique,* les machines sont de plus en plus singulières – mieux elles simulent et moins elles représentent. *Dans les discours,* l'ambition des chercheurs n'a pas évolué depuis les années cinquante : c'est la quête d'un modèle ou d'un principe de l'intelligence, de l'alphabet de l'esprit. Avec la conjecture formulée à l'occasion du séminaire de Dartmouth, les premiers chercheurs récusent d'avance la singularité de leur objet d'étude. Si toutes les caractéristiques de l'intelligence peuvent être décrites d'une manière si précise qu'une machine peut les simuler, c'est qu'*aucune* caractéristique de l'intelligence n'est singulière, elle ne présente pas d'aspect pour lequel « aucune comparaison n'est possible puisqu'il n'existe aucun comparant au moyen duquel se le figurer » – « ce en quoi il *n'y a rien de tel* ». Les entités intelligences partageraient toutes, en tant qu'elles sont intelligentes, les *mêmes* caractéristiques. Ce qui fait de nous des entités intelligentes serait commun, et représentable de bout en bout. L'ensemble des détails qui singularisent un esprit pourrait être mis de côté – « négligé » – pour ne garder que les traits communs qui constituent l'intelligence. Mais, en écartant ainsi la singularité, les auteurs de l'appel à projet de Dartmouth se privent de comprendre en quoi un esprit peut être original. Est-il tout de même possible de décrire et de reproduire les autres caractéristiques de l'intelligence, ou bien, en se privant de son originalité, se privent-ils de l'intelligence tout entière ? Autrement

dit, est-ce qu'une intelligence, pour être intelligente, doit être singulière ? L'originalité ne semble pas être une caractéristique parmi d'autres, que l'on pourrait ajouter à un esprit par ailleurs fonctionnel. Nous avons vu qu'il nous fallait inverser la perspective selon laquelle le hasard ou la créativité « interviennent » dans un cours des événements, par ailleurs déterminé. De la même manière, il nous faut inverser la perspective selon laquelle notre « originalité » serait seconde. Nous naissons comme un esprit singulier, puis nous formons, d'une manière singulière, des notions communes. En d'autres termes, si la singularité est première ce serait une erreur de vouloir fabriquer un esprit « commun », *puis* le rendre original ou créatif. L'originalité n'est pas une caractéristique que l'on peut ajouter, elle est première.

Les programmeurs se demandent *ce qu'il faudrait ajouter* – quel « module » de liberté ou bien quelle « part d'aléatoire » – pour qu'un programme puisse changer ses propres règles. Étant donné que certaines écoles qualifient un tel programme de *singularité*[850], on est amenés à se demander si ce n'est pas en effet ce qui « manque » aux machines pour penser, non pas au sens où l'entendent ces écoles (une « explosion de l'intelligence ») mais au sens que lui donne la tradition philosophique, c'est-à-dire une inscription dans un temps et dans un lieu qui confère un *point de vue* impossible à représenter. Ce qui fait un interlocuteur, c'est la singularité de son point de vue. Pour que je pense comme lui, il aurait fallu que je naisse dans le même corps et que je vive chaque événement de sa vie, aux mêmes endroits, de façon à développer les mêmes préférences, les mêmes habitudes, les mêmes principes.

La conjecture de Dartmouth dit la possibilité du zombie : nous pouvons décrire et reproduire chacune des caractéristiques de notre intelligence. Nous pouvons fabriquer une machine qui calcule, qui perçoit, qui parle. Mais pour que le zombie participe de l'esprit, pour le calcul ne soit pas aveugle et que quelqu'un perçoive, il faut un point de vue, autrement dit une singularité. Sans cela, nous en manquons la réalité. Cela implique que, pour « copier » l'intelligence, il faudrait la copier tout entière ou ne rien en recopier. Si seule une partie en est reproduite – une « fonction » ou un « module » –, sa singularité, et donc son point de vue, manquent. Les notions communes ou les concepts élémentaires ne sont pas le point de départ de l'intelligence. Ils en sont l'horizon, le bien commun qu'il nous faut fabriquer et, en chaque lieu et à chaque époque, recommencer.

850 Jean-Gabriel Ganascia, *Le mythe de la singularité, Faut-il craindre l'intelligence artificielle ?* Paris, Éditions du Seuil, 2017.

3.6.9. Pour demain

Avec le triomphe de l'école connexionniste sur l'école symbolique, le projet d'intelligence artificielle prend un virage empiriste. En pratique, le projet d'intelligence artificielle abandonne progressivement les grands principes et se rapproche de la singularité des phénomènes. Mais les succès pratiques rencontrés, au lieu de valider l'empirisme qui les permettent, ont servi au contraire à raviver les prétentions rationalistes des chercheurs. Forte de ses incroyables succès techniques, l'école connexionniste est aujourd'hui en mesure de relancer, avec un écho retentissant, la grande promesse du projet d'intelligence artificielle. Les programmes de *deep learning* n'ont pas « résolu le mystère » de l'intelligence, mais ils ont convaincu un grand nombre d'acteurs qu'ils sont sur le point d'y arriver. Pour Yann LeCun et ses collègues, « cela ne fait pas de doute » que la fabrication de machines intelligentes est au bout du chemin :

> Est-ce qu'on va être capables de construire des machines d'intelligence générale, qui sont supérieures à l'humain dans tous les domaines ? Et ça c'est pas pour demain, on a pas les technologies pour ça, on a même pas les concepts, on a pas les mathématiques, enfin il nous manque plein de choses pour ça. Alors on a des tas d'idées, on y travaille, ça pourra prendre peut-être cinq ans, dix ans, vingt ans, on sait pas – peut-être trente ! – avant de trouver les concepts puis ensuite ça prendra d'autres décennies pour arriver à faire fonctionner […]. Cela ne fait pas de doute pour la plupart des gens dans le domaine que c'est possible. On a pas du tout l'idée que l'intelligence humaine ait quelque chose de magique derrière […]. On pense qu'on peut très bien l'atteindre avec des ordinateurs disons relativement classiques[851].

Yann LeCun a l'honnêteté de souligner qu'il manque « les concepts ». Aucun chercheur n'est capable d'expliquer selon quels principes fonctionneront ces machines. La définition claire de *ce qu'est l'intelligence,* qui nous aurait permis de dissiper une partie du flou autour de la notion d'intelligence artificielle, est laissée en suspens, sans que cela n'entame leur confiance dans le projet d'intelligence artificielle. Puisqu'il n'y a rien « de magique derrière » l'intelligence humaine, ils finiront par en percer le mystère. Mais n'est qu'une fois qu'elle aura été reproduite par les machines que l'on connaîtra l'essence de cette intelligence que l'on cherche à reproduire. On sait que les machines à venir seront « d'intelligence générale », mais cette dernière ne sera définie qu'une fois les machines mises au point. On peut dire que le projet d'intelligence

851 Yann LeCun, « L'Intelligence artificielle dans nos têtes, avec Yann Lecun et Enki Bilal », *YouTube,* 30 janvier 2018, https://www.youtube.com/watch?v=ZjVQzkfRQ90 page consultée le 20 novembre 2020.

artificielle vise à « réaliser » l'intelligence, dans les deux sens du terme : la fabriquer (la « réaliser ») et la comprendre (« réaliser » ce qu'elle est).

Le *deep learning* n'a pas fourni les clefs de l'intelligence mais il s'en est approché d'un pas, et les chercheurs ont encore « des tas d'idées », notamment autour de l'apprentissage auto-supervisé[852]. L'intelligence obéit à des principes connaissables que l'école connexionniste serait sur le point de découvrir, et qui permettront aux machines d'acquérir une « intelligence générale » ou au moins une « intelligence de niveau humain[853] ». Les deux autres ténors du *deep learning,* Geoffrey Hinton et Yoshua Bengio, font preuve d'un optimisme similaire. Pour Yoshua Bengio, la bonne méthode pourrait passer par un couplage avec des travaux plus « symboliques », notamment autour de la compréhension du langage naturel[854]. Tandis que pour Geoffrey Hinton, il suffit de prolonger les voies ouvertes par le *deep learning,* qui est selon lui déjà un modèle cohérent de la pensée : chaque pensée correspond à la matrice des poids et des biais de nos neurones à un instant donnée – une pensée n'est rien d'autre qu'un « très gros vecteur[855] ».

Ce bel optimisme n'est pas nouveau. Dès la fondation de la discipline, les chercheurs en intelligence artificielle agacent leurs collègues par leurs certitudes et leurs promesses : « plusieurs choses portaient à croire dans les débuts de l'IA en un progrès rapide, qui incitèrent les premiers chercheurs à faire preuve de cet optimisme excessif devenu aujourd'hui la marque du domaine[856]. » Hans Moravec le rappelle dans un article de 2009 :

> Dans un contraste frappant avec l'explosion imprévue des ordinateurs dans notre quotidien, l'ensemble de l'initiative de la robotique a complètement échoué à réaliser les prévisions des années cinquante. À cette époque, les experts, fascinés par les capacités miraculeuses des ordinateurs, pensaient que si seulement les bons logiciels étaient écrits, les ordinateurs pourraient devenir les cerveaux artificiels de robots autonomes sophistiqués. En une ou

852 Kyle Wiggers, « Yann LeCun and Yoshua Bengio: Self-supervised learning is the key to human-level intelligence », *VentureBeat,* 2 mai 2020, https://venturebeat.com/2020/05/02/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/ page consultée le 2 mars 2021. L'article est un compte-rendu des interventions de LeCun et Bengio à la conférence ICLR 2020.

853 Depuis peu, Yann LeCun privilégie l'expression « intelligence de niveau humain » pour se démarquer de l'expression « AGI », artificial general intelligence, qui lui paraît exagérée.

854 Kyle Wiggers, *op. cit.*

855 Geoffrey Hinton, interviewé par Andrew Ng, « Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton », Youtube, 9 août 2017, https://www.youtube.com/watch?v=-eyhCTvrEtE, page consultée le 4 octobre 2019.

856 Daniel Crevier, *A la recherche de l'intelligence artificielle,* Champs Flammarion, 1999, p. 18.

> deux dizaines d'années, pensaient-ils, de tels robots viendraient laver nos sols, tondre nos pelouses et en général éliminer les corvées de nos vies[857].

Après avoir mentionné l'optimisme excessif des chercheurs des années cinquante[858], on pourrait s'attendre à ce que Moravec fasse preuve de plus de réserve. Au contraire, il l'utilise comme préambule à une réitération de *la même* promesse :

> Cependant, je suis convaincu que le vieux rêve d'un robot autonome, universel et utile sera réalisé dans un futur proche. En 2010 nous verrons des robots mobiles aussi grands que les humains mais dotés de capacités cognitives similaires, par bien des aspects, à celles d'un lézard. Les machines seront capables de réaliser des corvées simples comme passer l'aspirateur, dépoussiérer, livrer des colis et sortir les poubelles. Je crois que d'ici 2040 nous aurons enfin accompli le but originel de la robotique et un des piliers thématiques de la science-fiction : une machine capable de se déplacer librement avec les capacités intellectuelles d'un être humain[859].

Moravec a beau connaître la sombre histoire des prophéties dans le domaine de l'IA, cela ne l'empêche pas de réitérer la même promesse que ses prédécesseurs. Ils avaient tort, mais lui a raison. S'il était possible de l'interroger en 2040, sans doute en irait-il de même vis-à-vis de sa propre opinion : il avait tort en 2009 mais il penserait à nouveau avoir raison, annonçant encore que des robots aux « capacités intellectuelles d'un être humain » sont pour le lendemain.

Jonathan Wang et Brian Potter, du *Machine Intelligence Research Institute,* ont rassemblé 257 prédictions relatives à l'intelligence artificielle faites entre 1950 et 2012, dont 95 donnent une date à l'arrivée de l'intelligence artificielle dite forte[860]. En analysant cette base

---

857 « In stark contrast to the largely unanticipated explosion of computers into the mainstream, the entire endeavor of robotics has failed rather completely to live up to the predictions of the 1950s. In those days experts who were dazzled by the seemingly miraculous calculational ability of computers thought that if only the right software were written, computers could become the artificial brains of sophisticated autonomous robots. Within a decade or two, they believed, such robots would be cleaning our floors, mowing our lawns and, in general, eliminating drudgery from our lives. » Hans Moravec, « Rise of the Robots, By 2050 robot "brains" based on computers that execute 100 trillion instructions per second will start rivaling human intelligence », *Scientific American*, 01/02/2008, https://www.scientificamerican.com/article/rise-of-the-robots-2008-02/ consulté le 25 janvier 2019

858 Une fois de plus, on observe une dissociation entre l'objectif technique et l'ambition « scientifique », puisque aujourd'hui nous disposons de robots autonomes lavant les sols et tondant les pelouses, mais personne ne les voit comme « les cerveaux artificiels de robots autonomes sophistiqués »

859 « Nevertheless, I am convinced that the decades-old dream of a useful, general-purpose autonomous robot will be realized in the not too distant future. By 2010 we will see mobile robots as big as people but with cognitive abilities similar in many respects to those of a lizard. The machines will be capable of carrying out simple chores,such as vacuuming, dusting, delivering packages and taking out the garbage. By 2040, I believe, we will finally achieve the original goal of robotics and a thematic mainstay of science fiction: a freely moving machine with the intellectual capabilities of a human being. » *Ibid*

860 La base de donnée est disponible à ce lien https://aiimpacts.org/miri-ai-predictions-dataset/

de données, Stuart Armstrong et Kaj Sotala ont remarqué une tendance nette malgré la grande variabilité des prévisions : les experts évaluent « systématiquement » l'arrivée de l'intelligence artificielle « dans 15-25 ans », et cela tout au long de l'histoire de l'intelligence artificielle[861]. Autrement dit, « l'intelligence artificielle est perpétuellement dans quinze ou vingt-cinq ans[862] ». En 1950, les machines intelligentes étaient annoncées pour 1970. En 1970, elles étaient prévues pour 1990. En 1990, on évoquait les années 2010, et ainsi de suite. Aujourd'hui encore, Yann LeCun pense que « cela va probablement nécessiter quelque chose comme deux décennies pour qu'on arrive à des résultats vraiment spectaculaires. Je ne veux pas trop m'avancer en ce qui concerne les échéances, mais cette estimation n'est pas n'importe quoi[863]. »

L'intelligence artificielle est donc *pour demain,* mais *toujours pour demain.* « *Pour demain* » est compris au pied de la lettre, en faisant abstraction de l'inscription temporelle de l'énoncé. « Demain » devrait signifier « demain à partir d'aujourd'hui ». Ici, demain est devenu *toujours* demain : à chaque « printemps de l'IA », aussi bien dans les années 1960, que dans les années 1980 et dans les années 2010, l'intelligence artificielle adviendra « dans vingt ans ». A mesure que la recherche évolue, la prédiction reste la même, et il y a fort à parier que, dans vingt ans, l'intelligence artificielle sera encore prévue pour « dans vingt ans ».

Dans une scène célèbre de *Through the looking glass,* la reine affirme à Alice qu'il y a de la confiture hier ou demain (« jam to-morrow »), mais jamais aujourd'hui :

> – […] La règle en ceci est formelle : confiture demain et confiture hier – mais jamais confiture aujourd'hui.
>
> – On doit bien quelquefois arriver à confiture aujourd'hui, objecta Alice.

861 « Not only were expert predictions spread across a wide range and in strong disagreement with each other, but there was evidence that experts were systematically preferring a "15 to 25 years into the future" prediction. In this, they were indistinguishable from non-experts, and from past predictions that are known to have failed. » Stuart Armstrong and Kaj Sotala. « How we're predicting ai–or failing to. » In Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, editors, *Beyond AI: Artificial Dreams*, Pilsen, University of West Bohemia, 2012, p. 52–75. Les auteurs approfondissent leur travail dans Stuart Armstrong, Kaj Sotala, Seán S. Ó hÉigeartaigh. « The errors, insights and lessons of famous AI predictions – and what they mean for the future. » *Journal of Experimental & Theoretical Artificial Intelligence* 26:3, 2014, p. 317-342

862 « AI is perpetually fifteen to twenty-five years into the future ». *Ibid*

863 « Les progrès vont être continus dans les transports, la médecine, la recherche d'informations, la communication. A brève échéance, l'IA va sauver des vies ! Mais cela va probablement nécessiter quelque chose comme deux décennies pour qu'on arrive à des résultats vraiment spectaculaires. Je ne veux pas trop m'avancer en ce qui concerne les échéances, mais cette estimation n'est pas n'importe quoi. » Stanislas Dehaene, Yann Le Cun, Jacques Girardon, *La plus belle histoire de l'intelligence, Des origines aux neurones artificiels, vers une nouvelle étape de l'évolution,* Robert Laffont, Paris, 2018, p. 225.

> – Non, ce n'est pas possible, dit la Reine. C'est confiture tous les autres jours. Aujourd'hui, cela n'est pas tous les autres jours, voyez-vous bien[864].

L'expression « jam to-morrow » est devenue en anglais un synonyme de « pie in the sky », autrement dit de promesses en l'air, en particulier dans le domaine politique ou économique[865]. L'intelligence artificielle, comme la confiture de la reine, est toujours pour demain et jamais pour aujourd'hui. Dans *Logique du sens,* Gilles Deleuze attribue le caractère paradoxal des textes de Lewis Carroll à l'impossibilité de figurer le devenir :

> Dans *Alice* comme dans *De l'autre côté du miroir*, il s'agit d'une catégorie de choses très spéciales : les événements, les événements purs. Quand je dis « Alice grandit », je veux dire qu'elle devient plus grande qu'elle n'était. Mais par là-même aussi, elle devient plus petite qu'elle n'est maintenant. Bien sûr, ce n'est pas en même temps qu'elle est plus grande et plus petite. Mais c'est en même temps qu'elle le devient. Elle est plus grande maintenant, elle était plus petite auparavant. Mais c'est en même temps, du même coup, qu'on devient plus grand qu'on n'était, et qu'on se fait plus petit qu'on ne devient. Telle est la simultanéité d'un devenir dont le propre est d'esquiver le présent[866].

Le devenir est comme la confiture de la reine : on peut « en avoir » hier ou demain, mais jamais aujourd'hui car son « propre est d'esquiver le présent ». Il est possible de l'anticiper ou de le percevoir après-coup, mais il n'est jamais sous nos yeux. Le présent de la transformation (« l'événement » selon les termes de Deleuze), nous échappe.

Nous avons vu comment, du fait de leur incapacité à revenir sur leurs propres principes, les algorithmes ne sont pas en mesure de changer. Ils peuvent se calibrer, et se recalibrer pour peu qu'on mette leurs données à jour, mais ils ne peuvent traverser une crise, la métamorphose leur échappe. Il n'existe pas de « principe de tous les changements » qui permettrait d'inscrire dans un programme la formule lui permettant de s'adapter à la nouveauté d'une situation. Si, contrairement aux programmes, les autres entités du monde « s'adaptent » et participent spontanément à ses transformations, c'est qu'elles y sont inscrites depuis la singularité d'un point de vue et d'une époque. Il n'existe pas de principe de l'intelligence, pas de principe de la faculté de comprendre et de s'adapter, pas plus qu'il n'y a d'Idées ou de modèles éternels qui

---

864 Lewis Carroll, *De l'autre côté du miroir* in *Œuvres*, Gallimard, collection « La Pléiade », 1990, p. 302. En version originale : « 'The rule is, jam to-morrow and jam yesterday—but never jam to-day.'
'It *must* come sometimes to 'jam to-day,''Alice objected. 'No, it can't,' said the Queen. 'It's jam every *other* day: to-day isn't any *other* day, you know.' » Lewis Carrol, *Through the looking glass,* Londres, Penguin Books, 1994.

865 « jam tomorrow promise of future treats etc. that never materialize », Della Thompson, *The Oxford Dictionary of Current English*, Oxford, Oxford University Press, 1993, p. 474.

866 Gille Deleuze, *Logique du sens,* Éditions de Minuit, 2005, p. 9.

règlent les mutations du monde. Nos règles de pensées, nos concepts, ne sont pas le point de départ, le soubassement, ou la « substance » de notre intelligence, ils en sont le point d'arrivée, ou plutôt l'horizon, les infrastructures de pensées que, comme celles de notre vie matérielle, il faut, par un effort considérable, maintenir et cultiver, c'est-à-dire aussi, selon l'esprit du lieu et de l'époque, toujours recommencer pour comprendre ce qui a eu lieu hier ou aujourd'hui – et fabriquer demain.

CONCLUSION

Résumé de la thèse

Qu'est-ce que l'intelligence artificielle ?

Le point de départ de la thèse a été l'affirmation, énoncée par des chercheurs en intelligence artificielle, que les programmes dits d'apprentissage profond (*deep learning*), descendants des réseaux de neurones, faisaient preuve d'intuition. Nous nous sommes donnés pour tâche d'étudier la pertinence de cette affirmation.

Dans une première partie, nous avons cherché à délimiter ce qu'est le projet d'intelligence artificielle en nous appuyant sur un commentaire des textes fondateurs, tout particulièrement l'article de 1950 d'Alan Turing et l'appel à projets du séminaire de Dartmouth, où est formulée la conjecture selon laquelle tous les aspects de l'intelligence peuvent être décrits avec une précision telle qu'une machine peut les simuler ; ainsi que sur une histoire succincte du projet d'intelligence artificielle mettant l'accent sur la place des réseaux de neurones. Partant de l'invention de l'ordinateur, du moment cybernétique, et de l'article fondateur de McCulloch et Pitts, nous avons examiné les Perceptrons de Frank Rosenblatt, la controverse à leur sujet, la mise au ban des réseaux de neurones qui en résulte, puis la « revanche des neurones[867] » qui prend sa source dans les années 1980, avec formation du groupe PDP et l'invention de l'algorithme de rétro-propagation, et trouve son couronnement dans les années 2010 avec le

867 Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones, L'invention des machines inductives et la controverse de l'intelligence artificielle », *op. cit.*

succès technique et industriel des algorithmes dits d'apprentissage profond (*deep learning*). À la faveur de cette histoire, nous avons pris le temps de détailler les principes et le fonctionnement effectif de certains réseaux de neurones, notamment le Perceptron, la rétro-propagation du gradient, ainsi que les réseaux convolutionnels. Nous avons également décrit comment, et pour quels problèmes, l'industrie utilise ces programmes, ainsi que les autres programmes de la famille de l'apprentissage machine (*machine learning)*.

Puis, nous avons examiné ce que les programmes d'apprentissage profond proposent en tant que modèles de l'esprit. La « connaissance » y étant encodée par l'expérience, c'est une représentation empiriste de la pensée. Ce sont, pour Dominique Cardon et ses collègues, des « machines inductives[868] » – l'« induction » y étant entendue comme équivalent à la connaissance par expérience, et plus précisément la capacité à généraliser : un réseau de neurones acquiert des connaissances, ou « apprend » des « représentations », en détectant les caractéristiques communes aux exemples d'une même classe qui lui ont été soumis, de manière à pouvoir qualifier correctement de nouveaux exemples possédant tout ou partie des caractéristiques identifiées. Selon Cameron Buckner, en inventant un dispositif capable de définir et de reconnaître des classes avec assez de souplesse pour prendre en compte la diversité des individus et des situations, les créateurs des réseaux convolutionnels apportent une solution pertinente à une controverse historique au sujet du fonctionnement de l'induction.

Les réseaux de neurones convolutionnels sont aussi l'occasion de proposer un modèle mathématique de l'abstraction. Lorsqu'un réseau de neurones est calibré, les objets « perçus » sont encodés dans un espace vectoriel dont chaque dimension correspond à une caractéristique discriminante. Dans ce modèle, l'« abstraction » est une région de cet espace vectoriel, région obtenue en transformant l'espace de manière à mieux séparer les différentes collections d'objets. Une fois l'apprentissage effectué et les régions définies, les réseaux de neurones peuvent servir à fabriquer des images d'une classe en effectuant un rééchantillonnage des points depuis la région correspondante, d'une manière qui pourrait être qualifiée d'« imagination artificielle », puisque les images s'inspirent de la réalité sans correspondre à aucun objet réel. Par translation dans l'espace vectoriel, les réseaux de neurones peuvent également servir à transformer ces images à l'envi, en modifiant leurs caractéristiques ou leur style.

Un encodage similaire a lieu lorsque les réseaux traitent le langage. Chaque mot est représenté par un vecteur simple qui, au cours de l'apprentissage, est transformé en un vecteur plus petit qui compresse les relations de proximité avec les autres mots. C'est une manière de

868 *Ibid.*

mettre en pratique l'hypothèse structuraliste selon laquelle les mots sont définis par leurs relations de similarités et de différences entre eux. Un tel mécanisme conduit à encoder les relations analogiques entre les mots et illustre le principe structuraliste selon laquelle l'analogie serait le constituant fondamental du langage. Cependant, s'ils fournissent une meilleure « image du langage[869] », les réseaux de neurones ne nous éclairent pas sur la faculté de parler. Ils sont plus intéressants en tant qu'outil d'investigation que comme modèle de l'esprit.

Pour ce qui est de la perception, de l'abstraction et de l'imagination, les réseaux de neurones proposent des modèles qui, s'ils divergent de celui du cerveau, n'en fonctionnent pas moins remarquablement bien. Cela permettrait-il de les qualifier de machines intelligentes, voire intuitives ? Elles peuvent en donner l'apparence mais, alors que pour Turing l'apparence de l'intelligence pouvait suffire, les chercheurs contemporains ne sont pas du même avis, – comme en témoignent leurs efforts pour se démarquer de Sophia et autres projets de « fausse » intelligence artificielle, ainsi que de la longue tradition de prestidigitation dont a hérité le projet. Pour Yann LeCun et ses collègues, se contenter des apparences reviendrait à tromper le public. Les recherches en intelligence artificielle ne sont sérieuses que lorsqu'elles visent à « vraiment » connaître et dupliquer les principes de l'intelligence. Ils admettent que leurs machines sont loin d'être intelligentes, mais elles participent de la « vraie » intelligence artificielle parce qu'elles finiront par le devenir. Elles ne sont *pas encore* intelligentes, alors que Sophia ne l'est *pas du tout.*

Pendant longtemps, les chercheurs ont fait preuve d'une certaine prudence, définissant l'IA comme la « science qui consiste à faire faire aux machines ce que l'homme ferait moyennant une certaine intelligence ». Mais aujourd'hui, sans aller jusqu'à ceux qui parlent de fabriquer une « superintelligence », le but affiché du projet d'intelligence artificielle est de « percer les mystères » de cette intelligence et d'en doter les machines – le recours à des astuces qui en donnent l'apparence ne suffit pas. Comme le formulaient Shannon et McCarthy, une machine qui se contente de « chercher la réponse dans un dictionnaire[870] » passe à côté de ce que nous sentons être l'intelligence. Selon la distinction de Descartes, qui préfigure l'expérience de pensée de la chambre chinoise de Searle, les machines agissent « par la disposition de leurs organes » alors que nous agissons « par connaissance[871] ». Pour Descartes, les machines peuvent faire les mêmes tâches que nous, et mieux que nous, mais elles ne peuvent disposer de

869 Juan-Luis Gastaldi « Why Can Computers Understand Natural Language ?: The Structuralist Image of Language Behind Word Embeddings », *op. cit*.
870 Claude Shannon et John McCarthy (ed.), *Automata Studies, op. cit.*, p. VI.
871 René Descartes, *Discours de la méthode*, *op. cit.*, p. 54-55.

la généralité, voire de l'universalité, que donne la connaissance. Là où les chercheurs en intelligence artificielle se démarquent du point de vue de Descartes, c'est que pour eux la *res cogitans* n'est pas séparée de la *res extensa.* N'importe quels éléments de la *res extensa* peuvent devenir *res cogitans,* pour peu qu'ils soient agencés de la bonne manière, une manière que les réseaux de neurones seraient en train de préfigurer. Cela remet également en cause la distinction antique entre chose naturelle et chose artificielle. En agissant « par connaissance », une machine intelligente serait à l'origine de son mouvement, elle aurait sa propre « tendance naturelle au changement », qu'Aristote réserve pourtant aux choses naturelles[872]. L'attrait d'une telle perspective est puissant, et Aristote le relève dans un autre texte : des machines disposant de leur propre tendance au changement dispenseraient les humains de travailler[873]. Elles rouvrent la perspective d'un âge d'or où les éléments produisent d'eux-même ce dont nous avons besoin – un monde sans travail. Mais, avant même que les premiers ordinateurs ne soient conçus, Ada Lovelace mettait les points sur les i : les machines pourront faire tout ce qu'on leur demandera, sauf prendre des initiatives. Elles ne pourront pas être « à l'origine » de leurs actions[874] – ce qui est une autre manière de réaffirmer l'argument de Descartes : elles ne feront qu'agir en vertu de leur dispositif, et jamais par connaissance. Par conséquent, il arrivera toujours une situation où elles seront prises en défaut. La particularité de leur dispositif n'ayant pas l'universalité de la raison, il faudra donc toujours faire intervenir un humain – l'âge d'or des machines autonomes est impossible.

Cent ans plus tard, Turing répond à Lovelace en inversant la perspective : il ne s'agit pas de croire que les choses artificielles deviendront naturelles mais de voir que les choses naturelles sont, tout comme les choses artificielles, dépourvues de spontanéité. Nous nous croyons créatifs mais nous sommes seulement surpris, du fait de notre manque d'attention et de mémoire. Pour peu que nous en ayons les capacités cognitives, rien ne devrait nous surprendre, et nous devrions pouvoir calculer toutes les conséquences de nos idées. Pourtant, en 1938, le même Turing défendait l'idée d'une distinction entre l'intuition, émergence de « jugements spontanés », et *l'ingenuity,* désignant les raisonnements explicites et le calcul[875]. L'intuition, affirmait-il, ne peut pas être décrite en des termes mécaniques. L'intuition serait la part naturelle de la pensée, tandis que le calcul ou le raisonnement explicite serait sa part artificielle. En 1950, Turing semble avoir changé d'avis. Tout comme l'oignon n'est qu'une agrégation de de peaux,

872 Aristote, *Physique, op. cit.,* p. 58.
873 Aristote, *Politique, op. cit.*
874 Ada Lovelace, « Sketch of the analytical engine invented by Charles Babbage, by LF Menabrea, officer of the military engineers, with notes upon the memoir by the translator », *op. cit.*
875 Alan Turing, « Systems of Logic Based on Ordinals », *op. cit.*

écrit-il, l'esprit pourrait n'être qu'une superposition de fonctions mécaniques. Mais, dans sa manière de défendre l'idée d'une pensée purement mécanique, Turing en donne à voir tout l'inverse. « Le texte », écrit Bruno Latour « déploie tout un opéra » où « l'invention littéraire » et « l'imagination la plus débridée » se mêlent à « l'invention technologique, à la métaphysique et au formalisme logique[876] ». Il fait preuve d'une débauche de « jugements spontanés » qui tient plus de l'intuition que de l'*ingenuity.*

Depuis soixante-dix ans, le projet d'IA prétend « enlever les peaux » de l'oignon de l'esprit en proposant des descriptions mécaniques pour chacune de ses facultés. Mais chaque programme proposé ne dure qu'un temps. Au moment de leur invention, certains programmes passent pour la représentation d'une faculté de l'esprit. Ils sont alors qualifiés d'intelligence artificielle. Puis d'autres inventions viennent leur voler la vedette. Soit ils sombrent dans l'oubli, soit ils continuent à être utilisés, mais comme des outils informatiques communs, sans faire référence à l'intelligence qu'ils étaient censés émuler. Les choses ne se passent donc pas comme dans l'analogie de la peau de l'oignon. Turing laissait entendre que, lorsqu'une faculté a été émulée, il est possible de passer à la suivante et d'aller ainsi, de faculté en faculté, vers « la peau qui ne contient rien ». L'histoire de l'IA montre autre chose : aucune faculté n'a été considérée comme expliquée *une bonne fois pour toute* « en termes purement mécaniques ». Pour chaque faculté, les modèles proposés ont été constamment revus, moins en fonction de leur pertinence théorique qu'en fonction de leur efficacité pratique dans la réalisation de tâches industrielles. Par exemple, la « peau de l'oignon » correspondant à la vision artificielle n'a cessé de faire l'objet de théories concurrentes, et les résultats spectaculaires obtenus dans les années 2010 n'y ont pas mis fin. Comme nous l'avons vu avec l'exemple du Perceptron de Rosenblatt, les modèles proposés font toujours l'objet d'une « flexibilité interprétative » – d'autant que les chercheurs passent moins de temps à évaluer la ressemblance entre un système de reconnaissance d'image et la vision effectuée par le cerveau, qu'à travailler à l'amélioration du système pour reconnaître des images sur internet. L'histoire de l'intelligence artificielle n'est donc pas celle d'un épluchage progressif de l'oignon, mais plutôt d'un recommencement perpétuel de chaque épluchure. Cela tient à des raisons techniques – les machines peuvent toujours être améliorées –, mais aussi à des raisons théoriques. Chaque machine est censée représenter une partie de l'esprit. Or, dès qu'il s'agit d'une machine, ou plus précisément d'un algorithme, d'une suite finie d'étapes, nous nous retrouvons face à un dispositif où l'esprit est superflu. Le but d'un algorithme est de permettre l'effectuation d'une tâche sans avoir à y

876 Bruno Latour, *Chroniques d'un amateur de sciences, op. cit.*, p. 63-65.

penser, d'« évacuer la pensée du calcul[877] ». Aussi, décrire une faculté de l'esprit « en termes purement mécaniques », autrement dit trouver un algorithme capable d'effectuer une tâche cognitive, revient à montrer que cette tâche n'a pas besoin d'esprit pour être effectuée. Fabriquer une machine qui joue au go, ce n'est pas montrer que cette machine est intelligente, mais prouver qu'on a été assez astucieux pour inventer un moyen de jouer au go sans avoir à recourir à l'intelligence. En conséquence, le projet d'intelligence artificielle va de déceptions en déceptions. Un programme peut réussir à imiter une de nos facultés, mais une fois que l'on comprend *ce qui produit ses réponses,* il est discrédité. Aucune tâche particulière ne permet de faire passer une machine pour intelligente puisque toute tâche, quelle qu'elle soit, du moment qu'elle est effectuée par une machine, sera disqualifiée comme requérant l'intelligence. À mesure que les machines sont de plus en plus capables, le « mystère » de l'intelligence va donc en s'épaississant : il est possible de jouer aux échecs, de reconnaître des formes, de rédiger des pastiches littéraires, de faire des mathématiques *sans intelligence,* ou en tout cas sans comprendre. Les chercheurs se tirent de ce paradoxe en renvoyant les choses à l'avenir : ce n'est pas que les machines ne peuvent pas être intelligentes, c'est qu'elles ne sont *pas encore* intelligentes – disent-ils. En attendant, chaque nouvelle invention est, provisoirement, qualifiée d'intelligente, tant que son algorithme n'est pas clair pour tout le monde. Une fois le dispositif vulgarisé, la machine perd son qualificatif : le public ne « voit » pas « où » peut se loger l'intelligence puisqu'il connaît les étapes du dispositif et comprend que, conformément à la définition de ce qu'est un algorithme, chacune d'elle peut s'effectuer sans avoir à y penser. L'intelligence artificielle est donc un « signifiant flottant » : depuis soixante-dix ans, le terme désigne des objets qui changent à chaque génération. Comme la « vraie magie » telle qu'elle a été étudiée par l'anthropologie, l'efficacité de l'idée d'intelligence artificielle est indépendante de ses réalisations pratiques. Ces dernières sont vite obsolètes, et vite remplacées. Le projet d'intelligence artificielle tire sa force d'une croyance collective, la fabrication à venir de machines intelligentes et l'âge d'or qui en résultera.

Au vu de ce qui précède, nous proposons une définition double de l'intelligence artificielle. Premièrement, le terme désigne les machines et algorithmes fabriqués *dans le but* de fabriquer des machines intelligentes mais pour lesquels il serait exagéré d'affirmer qu'ils font preuve d'intelligence. Sous cet aspect, il s'agit d'un signifiant flottant puisque les machines et algorithmes désignés comme « de » l'intelligence artificielle ne cessent de changer. Deuxièmement, le terme désigne le rêve, toujours renvoyé à l'avenir, de machines intelligentes

877 Gérard Berry, cité par Philippe Flajolet, Étienne Parizot, « Qu'est-ce qu'un algorithme ? », interstices.fr, 2004.

qui mettront fin au travail. La notion d'intuition fait le lien entre les deux définitions en désignant *ce qui manque* aux machines inventées (première définition) pour qu'elles atteignent l'horizon du projet, les machines intelligentes, autrement dit pour qu'elles soient conformes à la deuxième définition.

Qu'est-ce que l'intuition ?

Dans la deuxième partie, nous avons cherché à délimiter ce qu'est l'intuition, en prenant pour point de départ les difficultés que posent sa définition. L'intuition se laisse difficilement décrire car son rôle est de s'effacer devant ce qu'elle nous donne à connaître. Considérer l'intuition indépendamment de ce à quoi elle nous donne accès est aussi ardu que d'entendre sa propre langue comme une langue étrangère. À cela s'ajoute sa variabilité – ce qui est évident pour l'un ne l'est pas forcément pour l'autre – et son inconstance – l'intuition d'un jour peut m'apparaître inepte le lendemain. Se fier sans réserve à l'intuition, c'est manquer de rigueur.

Etymologiquement, l'intuition renvoie à la vue. C'est une catachrèse : la pensée a des yeux, au même titre qu'une table a des pieds. D'autres comparaisons permettent de mieux cerner l'intuition : via celle de la lumière, l'accent est mis sur sa capacité à « éclairer » les choses, et son aspect commun, voire universel. Des comparaisons avec d'autres sens (le nez, le toucher, l'état du ventre) la dépeignent au contraire comme tributaire de la singularité d'un point de vue ou d'une situation. Elle est, dans tous, les cas, une perception, ce qui sous-entend que la pensée (ce qui est perçu par l'intuition), est un donné. Ce donné, s'il est parfois dit immatériel (*res cogitans*), n'en est pas moins décrit selon un ensemble de termes empruntés au monde matériel (*res extensa*). On invoque l'espace, les solides, les liquides (cohésion), l'architecture ou le corps dans une série de locutions qui s'appuyent sur la chose étendue pour décrire la chose pensante.

Un autre moyen d'approcher l'intuition est de la comparer avec ce à quoi elle est communément opposée : la déduction, et plus généralement la pensée qui suit des règles. L'intuition désigne les jugements spontanés, et s'oppose à la pensée construite, voire « aveugle », lorsque l'attention privilégie la conformité à des règles plutôt que le sens des propositions. Descartes et Leibniz s'affrontent au sujet du type de pensée qui garantit la certitude. Pour Descartes, c'est la contemplation attentive des idées, tandis que pour Leibniz, c'est la rigueur de leur construction qui permet d'arriver à la vérité. Ils convergent cependant

sur un point : lorsqu'il s'agit des principes, les règles ne permettent pas de s'orienter, et l'intuition doit se débrouiller seule. Les principes, et plus généralement les vérités primitives, fondements, axiomes, « idées claires et distinctes », et données brutes des sens, sont l'affaire de l'intuition. Autrement dit, l'intuition est le premier organe de la philosophie.

Par contraste, il est courant de dénigrer la pensée « aveugle » ou pensée formelle. Lorsqu'il effectue une procédure, un opérateur n'a pas besoin de comprendre ce qu'il fait. Il met de côté son esprit d'initiative et ses capacités critiques. Il ne considère les signes que sous leur versant matériel, sans s'attarder sur leur signification. Il met le sens de côté et se contente de respecter la syntaxe. Une grande partie des écoliers vit mal le statut que prend le langage, comme signe dépourvu de sens, et l'injonction qui leur est faite de transformer une suite de signes selon des règles sans se poser de questions. Leur capacité critique est vue comme un perturbateur potentiel, qui pourrait entraver le bon déroulement de la procédure. Ces caractéristiques de la pensée « aveugle » nous permettent de mieux cerner ce que fait la pensée lorsqu'elle « voit », autrement dit le rôle que joue l'intuition. L'intuition permet d'attribuer un sens à différents niveaux de l'opération : c'est « avoir en tête » la signification des signes manipulés (le sens comme dénotation), mais aussi le résultat de la procédure (le sens comme direction), et l'intérêt de sa réalisation (le sens comme finalité).

Il est exagéré de considérer que la pensée aveugle est une ablation complète de l'intuition. Cette dernière doit intervenir en amont, lorsqu'il s'agit de concevoir ou de choisir la procédure, et à la fin, lorsqu'il faut interpréter le résultat. Loin de remplacer l'intuition, elle est, pour Leibniz, son allié. Nous avons besoin de la rigueur du formalisme pour suppléer à nos fragilités : inconstance de l'intuition, faiblesse de notre attention, de notre entendement, de notre mémoire, etc. Elle permet de déléguer à des supports extérieurs le soin de préserver une cohérence à nos réflexions. Nous pouvons ainsi manipuler des objets trop complexes pour avoir toutes leurs caractéristiques à l'esprit. L'élaboration de règles communes sur la représentation et l'évolution des connaissances est indispensable pour s'affranchir de la pénible variabilité de l'intuition et garantir la communicabilité et le progrès des connaissances. Enfin, contrairement à ce que laisse penser sa réputation, la pensée aveugle est aussi un moyen d'inventer *(ars inveniendi*). Comme une canne donne la possibilité de tâtonner au-delà de ce que la vue perçoit, la pensée aveugle permet l'émergence de formes pertinentes inconnues, que l'intuition seule aurait été en peine d'imaginer. Mais, privée d'intuition, la pensée aveugle seule est-elle en mesure de produire une invention ? Nous retrouvons les termes du débat entre Lovelace, qui affirme, à l'occasion de son commentaire de la machine de Babbage, qu'un système formel mécanisé ne saurait être à l'initiative de ses idées, et Turing, qui lui réplique cent ans plus tard

que les humains ne sont pas plus créatifs que les machines : si les uns comme les autres nous surprennent, c'est parce que nous ignorons leurs mécanismes, et non parce qu'ils seraient capables d'invention – autrement dit, l'ensemble des opérations de l'esprit peut être décrite dans les termes de la pensée aveugle, y compris l'invention. Quelques années après l'article de Turing, la conjecture formulée à l'occasion du séminaire de Dartmouth réitère ce postulat : toutes les caractéristiques de l'intelligence, y compris l'intuition et l'invention, peuvent être simulées par une machine.

Déjà, Leibniz avait formulé le rêve d'étendre le formalisme des mathématiques aux autres domaines de la culture (métaphysique, morale, politique, droit, médecine). Comme les mathématiques, ces derniers reposent sur des systèmes de signes, mais sans avoir la même rigueur : les termes sont mal définis, les règles de démonstration ne sont pas explicites, etc. Leibniz imagine une forme de calcul qui pourrait s'y appliquer et permettrait de départager les controverses. La pensée aveugle ne serait plus réservée aux mathématiques mais apporterait ses bénéfices (clarté, fiabilité, communicabilité, etc.) aux autres domaines du savoir. Cependant, même en mathématiques, il n'est pas possible de se passer de l'intuition pour compter exclusivement sur la rigueur formelle. Au tournant du vingtième siècle, les mathématiciens entreprennent de fixer les définitions, les axiomes et les règles qui structurent leur discipline. Le projet formaliste, porté par Hilbert, fait miroiter la perspective de mathématiques entièrement définies et réglées, où les théorèmes sont produits mécaniquement. L'intuition garde toutefois un rôle, pour tout ce qui est de l'ordre des métamathématiques, autrement dit lorsqu'il s'agit de faire évoluer les règles, les définitions et les axiomes des mathématiques. L'intuition n'est pas évacuée, elle est seulement réservée au travail sur les fondements. À l'inverse, certains mathématiciens et logiciens, comme Poincaré et Gödel, témoignent du rôle central que l'intuition joue dans leur manière de travailler, voire, comme le fait Brouwer, considèrent qu'elle fonde les mathématiques. Pour les intuitionnistes, un énoncé mathématique n'est valide que si un esprit peut le construire, autrement dit en faire l'expérience.

Les supports de calcul ont évolué selon plusieurs tendances qui se sont renforcées mutuellement (déléguer les opérations de l'intellect vers le sensible assisté par un support matériel, simplifier les opérations à effectuer, affranchir le support matériel des interventions humaines, et généraliser les opérations pour que le moins de machines possibles puissent effectuer le plus de calcul) et trouvent leur point culminant dans les travaux de Turing. En 1936, il décrit une machine capable de calculer tout ce qui est calculable, en définissant ces notions grâce à seulement deux signes (0 et 1) et trois opérations. Cette machine est universelle : un seul modèle de machine peut exécuter n'importe quel algorithme. Tout comme la

« caractéristique » de Leibniz, la machine de Turing se base sur un atomisme, une réduction de tous les termes à des primitives binaires assemblées selon des règles simples de composition. À la différence de la caractéristique, elle ne formalise pas les concepts mais les procédures. Par contre, elle est effectivement universelle : tous les algorithmes sans exception peuvent être formalisés de cette façon. Avec l'invention de l'ordinateur, ce qui est une universalité théorique ouvre la possibilité d'une universalité pratique : si la machine de Turing peut effectuer tous les algorithmes, peut-être l'ordinateur pourra-t-il effectuer toutes les procédures humaines ? Pour le philosophe de Göttingen, il s'agissait d'appliquer le formalisme des mathématiques à la pensée non mathématique. En 1950, pour Alan Turing, il s'agit plutôt de se demander si, à force de raffinement, la prothèse pourrait se substituer à l'organe qu'elle assiste, autrement dit si un ordinateur pourrait penser. Avec la conjecture de Dartmouth et la formation de l'école symbolique, les premiers chercheurs en intelligence artificielle adoptent cet horizon de recherche. Pour eux, toutes les opérations de l'esprit peuvent se décrire sous la forme d'une procédure, autrement dit comme de la pensée aveugle. Ce que nous appelons « intuition » ne désigne que les aspects de la pensée dont nous ignorons le fonctionnement. La « pensée aveugle » ne correspond pas à un moment du raisonnement qu'il faudrait suivre pour s'assurer une rigueur adéquate et provoquer de nouvelles découvertes, c'est *toute la pensée* qui est, en deçà des apparences, une manipulation de symboles selon des règles. Avec le déclin de l'école symbolique et son remplacement par l'école connexionniste, cette approche trouve quelques nuances, mais le postulat fondamental demeure identique. Dans les deux cas, l'intuition n'est pas considérée comme une part non mécanique de l'esprit, mais comme une part mécanique dont le fonctionnement ne nous est pas encore connu. De ce point de vue, l'intuition apparaît comme le « reste » inexpliqué d'une pensée tout entière descriptible de manière formelle. Pourtant, notre expérience montre l'inverse : nous produisons des jugements spontanés, que nous formalisons *a posteriori* pour pouvoir les vérifier, les comparer, et les communiquer. Historiquement, l'intuition précède le formalisme. Les mathématiques sont nées et ont pu produire quantité de résultats pertinents bien avant les efforts de formalisation des dix-septième et dix-neuvième siècles.

Avec les théorèmes d'indécidabilité, de nombreux de travaux, dont ceux de Gödel lui-même, ont essayé d'établir une preuve de la nécessité de l'intuition. Mais il apparaît que les limites du formalisme ne permettent pas d'extrapoler avec certitude de la nécessité de l'intuition. Tout se passe comme si nous ne pouvions faire mieux que de partager l'intuition de la nécessité de l'intuition, sans que cette intuition ne puisse trouver de confirmation par le raisonnement explicite.

La pensée symbolique est une prothèse qui permet de remédier à la fragilité de l'intuition. Cependant, l'efficacité de la prothèse n'implique pas pour autant qu'on puisse se débarrasser de l'intuition. Croire qu'il est possible de se passer de l'intuition et réduire la pensée à son formalisme reviendrait à s'arracher les yeux sous prétexte que l'on dispose d'excellentes lunettes. Nous ne pouvons pas priviligier un type de pensée plutôt que l'autre : l'intuition prend en charge l'origine des idées, leur nouveauté, leur sens, tandis que la formalisation leur confère rigueur et communicabilité. Négliger la forme, c'est perdre en précision, en lisibilité, en vérifiabilité. Négliger l'intuition, c'est porter un discours sans prendre en compte les hypothèses et les vérités premières qui le soutiennent.

Le parcours effectué nous permet de donner une définition de l'intuition : elle désigne les opérations de l'esprit qui ne s'expliquent pas « en termes purement mécaniques », autrement dit, selon l'expression de Turing, l'« esprit réel[878] ». Comme la matière ou le temps, comme la rose de Silesius, elle existe « sans pourquoi[879] ». L'intuition désigne la part de la raison qui est sans raison : « vérités primitives, principes, fondements, axiomes, 'idées claires et distinctes' ou données brutes des sens, selon le goût philosophique de chacun[880] ». Nous en faisons l'expérience lorsque, comme le décrit Poincaré, les « idées surgiss[ent] en foule[881] ». Elle est l'origine des formes, tout ce qui vient à l'esprit et dont les formes constituées (images, sons, énoncés, algorithmes…) sont des traces plus ou moins pérennes – selon que nous les inscrivons, ou non, sur des supports. Ce surgissement est spontané, mais il a besoin de nous, il s'effectue à la faveur de notre attention. La conscience que nous en avons forme le terreau sur lequel la « culture » peut se déployer. Notre interprétation des formes les fait résonner avec d'autres formes et leur permet de croître et de se perpétuer. En d'autres termes, nous les concevons : nous les comprenons, nous les engendrons, nous en ressentons du plaisir. Cette complicité avec les formes constitue la vie de l'esprit ou le « naturel » de la culture au sens étymologique de « nature » : sa naissance et sa croissance spontanée.

La même intuition préside au surgissement de nouvelles formes et à l'interprétation de formes déjà constituées. « Comprendre une idée » requiert la même intuition qu'« avoir une idée ». Dans les deux cas l'intuition est *partiellement involontaire*, (selon le mot de Nietzche, « une pensée vient quand 'elle' veut et non pas quand 'je' veux[882] »), bien qu'elle exige souvent, comme le raconte Poincaré, un effort conséquent – nous « faisons » une idée comme on « fait

878 Alan Turing, « Les ordinateurs et l'intelligence », *op. cit.*, p. 167.
879 Martin Heidegger, *Le principe de raison, op. cit.*, p. 102-109.
880 David Rabouin, « Penser comme un pied », *op. cit.*
881 Henri Poincaré, *Science et Méthode, op. cit.*, p. 50-51.
882 Friedrich Nietzsche, *Par-delà bien et mal, op. cit.*, p. 640.

ses dents » ; *contingente*, elle peut avoir lieu, ou non, sans raison particulière ; *éphémère*, ses traces perdurent et permettent de la renouveler, mais leur intuition est évanescente ; et *singulière,* elle est conditionnée par la personne, le lieu, le moment. Expérience singulière, l'intuition est aussi expérience de la singularité des formes et des situations. Elle nous permet d'appréhender en quoi elles ne sont comparables à rien d'autre : leur caractère exceptionnel.

L'expérience du sens, autrement dit l'interprétation d'une forme ou d'une situation, se décline de nombreuses manières : il s'agit de percevoir ce à quoi elle renvoie en tant que signe (dénotation), ce qu'elle annonce (visée), ce qu'elle sert (finalité) et ce sur quoi elle se fonde (principes). La perception des principes est l'aspect le plus caractéristique de l'intuition, puisqu'il s'agit de percevoir ce qui fonde une forme ou une situation, autrement dit leur part originaire (sans raison). Cela requiert un mode de raisonnement particulier où la pensée *revient sur ses pas* vers des notions fondamentales comme le temps, la matière, la vie, mais aussi l'intuition et ses notions voisines (le sens, la compréhension). Chacun fait l'expérience de ces concepts, en a une idée confuse, mais peine à les définir. L'intuition nous permet de manipuler ces notions si difficiles à concevoir mais sur lesquelles tout le reste de l'édifice du savoir repose. Malgré ses inconvénients majeurs – difficultés à la décrire, instabilité, variabilité interpersonnelle, absence de fondement – nous ne pouvons pas faire l'impasse sur l'intuition, elle est le fond sans fondement à partir duquel tout le reste de la pensée naît.

L'intuition est ce qui permet de voir le sens d'une forme, d'une situation, mais aussi d'une opération. Elle est présente lorsque nous pensons à ce que nous faisons, et peut s'absenter aussi bien de nos actions manuelles que de nos opérations cognitives : nous pouvons penser à quelque chose *sans y penser* – sans penser au fait que nous y pensons. C'est le cas de la « pensée aveugle », lorsque nous appliquons un algorithme sans prendre en compte le sens de l'opération, autrement dit lorsque nous suivons un enchaînement d'arguments (sa « trame démonstrative », selon les termes de Michel Foucault[883]) sans en effectuer l'interprétation pour nous (sa « trame ascétique »). Ainsi, l'« esprit réel » ne se trouve pas en mettant de côté toutes les opérations mécanisables de l'esprit, mais au sein de chacune d'entre elles, selon que nous l'effectuons, ou non, en y pensant, et en faisant, ou non, l'expérience du sens.

Le modèle de l'intuition proposé par l'apprentissage profond (*deep learning*) ne restitue qu'une infime partie de ces caractéristiques. L'« intuition » qu'il met en œuvre est un circuit de stockage et d'activation d'automatismes acquis par l'expérience, qui permet d'obtenir des réponses sans savoir comment elles se sont formées. Le modèle illustre comment, à partir de

883 Michel Foucault, « Mon corps, ce papier, ce feu », *op. cit.*, p. 258.

données sensorielles, il est possible de différencier, de classer et de manipuler les objets du monde, et comment un processus d'essai-erreur peut aboutir à la formation d'automatismes qui fournissent de bonnes réponses sans que nous sachions d'où elles viennent – c'est l'« opacité » des réseaux de neurones. Mais une telle conception brouille la distinction entre intuition et pensée aveugle, elle laisse de côté nos capacités à comprendre et à formuler des conjectures : elle rend compte de notre capacité à fournir des solutions, sans savoir d'où elles viennent, dans le cas de problèmes déjà connus, mais pas pour des problèmes nouveaux. Elle échoue à restituer notre faculté à *concevoir,* c'est-à-dire à la fois à comprendre les idées et à en former de nouvelles – à les *interpréter*.

De ce point de vue, la conjecture formulée à l'occasion du séminaire de Dartmouth, selon laquelle tous les aspects de l'apprentissage ou toute autre caractéristique de l'intelligence peuvent en principe être décrits d'une manière si précise qu'une machine peut être fabriquée pour les simuler, est à la fois juste et fausse. Juste, parce que toutes les opérations qu'effectue notre esprit pourraient se décrire de manière mécanique ; et fausse, parce qu'en les décrivant ainsi, en manquant l'intuition qui les fait naître et devenir, nous manquons également ce qui nous amène à les percevoir comme intelligentes : le sens et la compréhension.

Les inventions du projet d'intelligence artificielle permettent de cultiver les formes (images, textes, sons) comme on cultive les virus en laboratoires, c'est-à-dire les mélanger, les faire évoluer et muter sans qu'un esprit « hôte » n'ait à les concevoir. Dans ce cas, l'argument de Lovelace semble ne plus tenir : peut-on dire que les machines sont « à l'origine » d'idées, d'interprétations, ou de nouvelles formes en général ? C'est ce que laissent penser les dernières prouesses de programmes comme Dall-E, qui lorsqu'on leur soumet quelques mots, proposent des images inédites les illustrant, ou encore AlphaGo. Le coup 37 du deuxième match était « magnifique et créatif » déclare Lee Sedol après avoir été vaincu[884], alors qu'il est lui-même considéré comme un des joueurs les plus créatifs de sa génération. Lee Seedol a-t-il seulement été *surpris* par AlphaGo, comme Turing disait être surpris par ses machines, ou bien peut-on dire du programme qu'il a été « à l'origine » de ce coup, et donc effectivement « créatif » ?

884 Greg Kohs, *AlphaGo, op. cit.*

Mécanisme et création

Dans une troisième partie, nous nous sommes penchés sur la question de la création, qui nous semble être au coeur du problème de l'intuition, et plus généralement de ce qui manquerait aux machines pour accéder à l'intelligence : que signifie être « à l'origine » d'une idée, et plus généralement d'une forme ? Les machines peuvent-elles être « à l'origine » de formes ? Avant d'aborder le sujet des machines créatives, nous nous sommes demandé si la créativité pouvait, voire devait, se passer de subjectivité, en examinant des cas d'inventions et de découvertes ayant eu lieu à la faveur d'une suspension de l'effort conscient. Dans les exemples abordés, l'effort subjectif conscient semble entraver le processus de création. Tout se passe comme si les inventeurs devaient laisser la place à un « autre en soi » pour qu'émergent des formes qu'ils n'auraient pas été capables, ou qu'ils s'interdisent, de concevoir. Demis Hassabis explique ainsi la plus grande efficacité d'AlphaZéro par rapport à son prédécesseur AlphaGo[885]. Alors que ce dernier s'appuye sur l'historique des comportements humains, AlphaZéro s'entraîne uniquement à l'aide de parties virtuelles. Affranchi des préjugés humains, AlphaZéro donnerait lieu à une plus grande exploration combinatoire, et serait donc un programme encore plus « créatif » qu'AlphaGo – pourtant déjà qualifié de créatif par Lee Sedol.

La « créativité » de programmes comme AlphaGo et AlphaZéro semble pourtant se distinguer de celle dont font preuve les humains. Pour tenter de mieux délimiter cette différence, nous revenons sur la notion de jeu, et sur la distinction proposée par Roger Caillois, qui oppose le *ludus*, jeu encadré par des règles, à la *paidia* ou « fantaisie sans règle ». Les programmes peuvent « jouer », avec une efficacité sans précédent, à un *ludus* mais non à une *paidia*. Cette dernière implique de pouvoir changer les règles et implique une réflexivité (à quelles règles se tenir quand nous changeons les règles ?) ou plus généralement une normativité (faculté à concevoir de nouvelles règles) qu'il est contradictoire d'attribuer à un programme mais que les humains – et le vivant en général – manifestent en situation de crise.

Dans ses textes sur l'invention scientifique, Gaston Bachelard offre une distinction analogue. Il identifie deux modes de pensée : l'un qui consiste à explorer l'espace des possibilités offert par un jeu d'axiome, et l'autre, qu'il qualifie de « révolutionnaire », consistant à revenir sur les axiomes pour les modifier, pour lequel il donne l'exemple de l'invention de la géométrie non euclidienne par Lobatchevsky. Il y aurait une raison

885 Sarah Knapton, « AlphaGo Zero: Google DeepMind supercomputer learns 3,000 years of human knowledge in 40 days », *op. cit.*

« prudente », qui n'avance que de manière incrémentale, en capitalisant sur ce qui existe déjà, et une raison « créatrice » qui « avance » en revenant sur ses pas, en remettant en question les savoirs préétablis. Pour Bachelard, il ne s'agit pas de privilégier l'une sur l'autre mais de conjuguer les deux.

Si le mouvement incrémental de la raison peut être formalisé et délégué à une machine, son mouvement réflexif échappe à toute formalisation. Comme le montre Bateson dans ses « métalogues », il n'est pas possible de fixer les règles qui président au changement des règles – ce qui est une autre manière de dire qu'aucun *ludus* ne peut rendre compte de la *paidia.* Voilà une nouvelle objection à la réponse que faisait Turing à l'argument de Lovelace : les machines ne nous surprennent pas de la même manière que les humains. Les machines peuvent jouer à des jeux définis (*ludus*) et faire émerger, grâce à leur capacité exploratoire, des combinaisons inconnues – et ainsi l'algèbre au pu être considérée, au dix-septième siècle, comme un *ars inveniendi*. Mais les humains peuvent varier leur interprétation du jeu, et surtout en modifier les règles. À la raison prudente qui permet d'« appliquer » un jeu fixe (*ludus*), ils combinent la raison créatrice, sans que l'une l'emporte sur l'autre. Si elle est seule, la raison créatrice manque de consistance et ne peut durer, elle tend vers l'internel. Tandis que sans raison créatrice, la raison prudente seule se laisse déborder par le cours des événements, elle tend vers l'éternel et perd prise avec son époque – en situation de crise, les algorithmes sont dépassés. Autrement dit, privilégier le *ludus* ou la *paidia* amène, dans les deux cas, à être à côté de l'histoire. Les participants aux « métalogues » de Bateson ne disent pas autre chose : à trop suivre les règles (excès de politesse), ou à n'en suivre aucune (« embrouillaminis »), une conversation manque le sens.

Dans des termes qui pourraient rappeler ceux de Bachelard, les auteurs de l'appel à projets de Dartmouth opèrent une distinction entre « la pensée compétente sans imagination » et « la pensée créative ». La différence, écrivent-ils, repose dans « l'injection de hasard ». Mais l'injection de hasard ne suffit pas pour garantir la *pertinence* des formes proposées. Aussi sont-ils contraints d'en appeler à une intuition qui guide et limite le hasard : le hasard n'intervient qu'en des lieux déterminés à l'avance, il peut proposer certaines possibilités mais pas d'autres, et ce qu'il produit est « guidé » par l'intuition. Ce hasard est similaire à celui qu'introduit un jet de dés dans un jeu de plateau : on sait ce qu'il va déterminer (le nombre de cases dont va avancer le pion), dans quelles proportions (entre 1 et 6) et ce qu'il en adviendra (une évolution du jeu avec un gagnant et des perdants). Un tel concept de hasard échoue à rendre compte de la créativité. Les auteurs de l'appel à projets de Dartmouth, et plus particulièrement Rochester,

sont les premiers à l'admettre. Le problème réside dans le fait que, comme le montre Fodor[886], l'invention peut faire feu de tout bois : n'importe quel élément du savoir peut se révéler pertinent. Pour en rendre compte, il faudrait donc un hasard « au carré », similaire à celui que Borges met en scène dans « La loterie à Babylone[887] », dont le lieu d'application, tout comme les effets et la fréquence, sont *aussi* dus au hasard. Mais comment programmer, ou même définir, un tel dispositif ? À la fin de la nouvelle de Borges, plus personne ne connaît les règles de la loterie. Elle est régie par une société secrète, si secrète qu'on en vient à douter de son existence.

Les auteurs de l'appel à projets de Dartmouth sont contraints de postuler un hasard circonscrit parce qu'il serait impossible de représenter, et donc de programmer, un hasard capable d'intervenir, d'une façon non prédéfinie, à n'importe quel moment du processus. Pour Clément Rosset, à partir du moment où l'on postule que les événements se déroulent selon un cours nécessaire où le hasard ne fait qu' « intervenir » de manière délimitée, on s'interdit de comprendre le hasard. Le véritable hasard ne se conçoit pas, « à partir de l'ordre ». C'est « un hasard d'avant la nécessité[888]. »

Si les auteurs de la proposition de Dartmouth formulent leur hypothèse selon le paradigme de l'intervention (intervention d'une dose de contingence dans une pensée « par ailleurs ordonnée »), c'est qu'ils ne peuvent pas faire autrement : leur point de départ est une description de l'intelligence à laquelle « manque » une faculté : la créativité. Ce « manque » est interprété comme un « manque » de hasard qu'il faut par conséquent « ajouter » au modèle. Or un hasard « ajouté » ne peut constituer un véritable hasard puisqu'il y a « intervention ». Autrement dit, aucun « manque » de hasard ne saurait être « comblé » par l'ajout de quelque chose d'autre (d'une « dose » de hasard). Prendre le hasard au sérieux, c'est admettre que celui-ci est premier. Les régularités, les structures, les stabilités transitoires que nous prenons pour des « natures » naissent du hasard – de la même manière que, selon Roger Caillois, le *ludus* procède de la *paidia.* Si pensée créative et pensée incrémentale cohabitent sans que l'une puisse exister sans l'autre, il y a tout de même prééminence de la pensée créative sur la pensée incrémentale. Il convient *d'inverser* le point de vue des auteurs de la proposition de Dartmouth : ce n'est pas le hasard (assimilé à la créativité) qui intervient au sein de l'ordre, mais différents ordres qui se constituent et se succèdent par hasard. Il n'y a pas « un hasard contrôlé dans une pensée par ailleurs ordonnée » mais émergence d'automatismes depuis un fond sans nécessité.

886 Jerry Fodor, *La modularité de l'esprit : essai sur la psychologie des facultés, op. cit.,* quatrième partie.
887 Jorge Luis Borges, « La loterie à Babylone », *op. cit.*
888 Clément Rosset, *Logique du pire, op. cit.*, p. 72.

Dès lors, la créativité n'est pas un privilège de l'humain ou du vivant mais l'expression de « l'aptitude de la matière à s'organiser spontanément[889] ». Nous rejoignons Turing, et les partisans du projet d'intelligence artificielle, dans leur volonté de remettre en cause un privilège spirituel de l'humain par rapport à la matière, mais pour des raisons inverses. Au lieu de considérer que l'humain relève de la même nécessité mécanique que la matière, nous postulons que l'un comme l'autre sont des structures éphémères qui naissent du même hasard. En conséquence, l'activité de l'artiste ou de l'inventeur ne consiste donc pas user d'un hasard « maîtrisé » par une intuition qui le guiderait.

À travers une analyse de l'alcoolisme, Bateson entreprend une critique de cette notion de maîtrise. L'alcoolique est victime de l'« épistémologie de la maîtrise de soi[890] » selon laquelle il devrait être « capitaine de son âme[891] » et le conduit à se comporter comme s'il pouvait maîtriser sa relation à la bouteille. Plus il met sa « maîtrise de soi » à l'épreuve et plus il sombre dans l'alcoolisme. La notion de maîtrise de soi ne résiste pas à l'épreuve, ni à l'observation : à mesure que j'observe mon environnement, mon comportement, mes pensées de façon à délimiter *sur quelle zone* s'exerce ma maîtrise, je verrai une réduction progressive du « soi » que je suis censé maîtriser : jusqu'à mon « intériorité », les pensées de mon « for intérieur », résistent à une description en termes de contrôle, tout particulièrement aux moments où j'en ai besoin, c'est-à-dire lorsque le contrôle est mis à l'épreuve. L'alcoolique doit donc, pour aller mieux, abandonner le principe de la maîtrise de soi, sans que cela puisse avoir lieu volontairement. Cela se fait à la faveur d'un épisode traumatique, d'une rencontre ou d'un événement contingent, que les alcooliques anonymes appellent « toucher le fond ».

Ce n'est donc pas le « soi », un hypothétique « gouverneur » logé dans son for intérieur, qui fait la conduite d'une personne, mais la collection d'automatismes incorporés au cours de son existence lui permettant d'évoluer avec plus ou moins d'aisance dans différents contextes. La notion d'habitus, élaborée par Pierre Bourdieu, nous permet de compléter le tableau ébauché par Bateson. Nous n'agissons pas en fonction d'un système de règles, logé dans notre tête, qui régirait nos actions, mais d'un ensemble d'inclinations – « un système de dispositions durables et transposables[892] » – qui se fait et se défait au fil de l'expérience. Contrairement aux notions de règles et de prémisses, l'idée d'un ensemble de dispositions incorporées et mobilisées par le contexte laisse plus de latitude à la cohabitation de dispositions disparates. S'il semble

889 « Hasard est précisément le nom qui désigne l'aptitude de la matière à s'organiser spontanément », Clément Rosset, *op. cit.*, p. 85.
890 *Ibid*, p. 285.
891 *Ibid*, p. 268.
892 Pierre Bourdieu, *Esquisse d'une théorie de la pratique, op. cit.,* p. 178.

déterminer notre conduite, l'habitus résiste à la comparaison avec un programme car il peut intégrer des automatismes contradictoires (habitus clivé) et surtout, il évolue. Il a une histoire qui suit un déroulement non programmé.

Nous nous appuyons ensuite sur François Roustang pour montrer que notre obstination à penser le sujet comme une entité fixe qui contrôle le corps est une forme de narcissisme. Malgré nos changements incessants, nous aimerions nous contempler dans une forme fixe, dont la machine est l'idéal. Mais la comparaison avec les machines nous persuade que nous dysfonctionnons et nous aveugle quant à notre capacité à changer, voire entrave celle-ci. Le propos de Roustang s'inscrit dans le cadre d'une critique de la théorie freudienne, à qui il reproche d'avoir délaissé la notion d'influence au profit d'une idée du sujet comme machine, plus à même de satisfaire à un certain idéal de scientificité, et comme machine *close*, plus à même de satisfaire un certain narcissisime. L'obstination moderne pour l'idée d'un sujet sans liens (« absolu ») n'est que l'effet d'un attachement contradictoire à une image de soi sans attachements. Il convient d'en revenir à une idée du sujet défini, et tenu, par ses relations. Mais cette dernière a l'inconvénient, du point de vue scientifique, de brouiller les frontières du sujet et du monde, et de produire, chez le narcissique, une réaction épidermique : s'il n'est pas le maître, c'est que l'autre terme de la relation est aux commandes. La conception moderne du sujet condense toute la puissance d'agir d'un côté ou de l'autre de la relation (paradigme du contrôle) au lieu de la répartir entre les deux (paradigme de l'influence ou de l'attachement) – comme en témoigne l'obsession contemporaine pour les machines qui « prennent le contrôle ». Pour que la machine puisse se dérouler dans le temps, autrement dit perdurer, il faut qu'elle entre en relation avec l'électricité, avec le programmeur, avec les matériaux qui la composent et qu'il faut remplacer. Ce sont autant de relations, et d'enchevêtrements, qui sont autant d'occasions de surprises et de détournements – la machine ne fait jamais exactement ce que lui demande celui qui la « contrôle » – et qui remettent en question l'attribution de l'action : elle n'est pas le fait d'un sujet solipsiste mais d'un collectif fragile, un ensemble d'alliances qui ne cesse d'être remis en jeu, chaque alliance demandant un travail de traduction et de médiation qui dévie le cours de l'action et l'amène à être « toujours légèrement dépassée par ce sur quoi elle agit[893] ». Les effets apportent toujours plus que les causes, et il y a du nouveau. Le sujet n'est pas « libre » au sens moderne mais mû par un assemblage hétéroclite d'attachements, assemblage perpétuellement renouvelé qui, comme dans la chimère de Baudelaire, nous remet sans cesse en mouvement sans qu'il y ait pour autant de direction – l'expérience de ce flux

893 Bruno Latour, *L'espoir de Pandore, op. cit.,* p. 318.

incessant et insaisissable (le fait que « ça aille ») nous procurant, selon les circonstances, un sentiment d'enfermement (la nausée) ou une joie sans raison.

Nous ne maîtrisons pas ce jeu de l'existence, mais nous aspirons tout de même, lorsqu'une situation change, à ajuster notre activité de la manière qui convient. En cas de crise, nous avons la capacité de reconsidérer nos principes pour laisser émerger de nouvelles règles d'action. Comment décrire ce changement de principes qui ne semble pas avoir de principe ? Surtout, comment faire la distinction entre un changement pertinent et un changement qui ne convient pas ? Selon Heidegger, la controverse entre Leibniz et Descartes peut se lire comme un désaccord sur les principes qu'il faut conserver et les principes qu'il faut suspendre. Heidegger prend parti pour Leibniz, contre Descartes, car ce dernier aurait manqué de *paideia.* Ce mot, à distinguer de la *paidia,* Heidegger renonce à le traduire et le définit comme « le don de discernement entre, ce qui face à des situations simples, est approprié et ce qui ne l'est pas[894]. » La notion de *paideia* pourrait nous éclairer dans notre recherche de ce qui fait la pertinence de l'invention, pertinence que la notion de hasard ne permet pas de restituer. Proposer de nouvelles combinaisons de formes – et les machines peuvent en proposer à foison –, est une chose, mais s'en est une autre de proposer une combinaison *intéressante.* Malheureusement, la *paideia* ne se laisse approcher que de manière négative : nous sentons lorsqu'elle vient à manquer, sans arriver à mettre le doigt sur ce qui, précisément, manque. La notion est introduite par Aristote, en ces termes, au livre IV de la *Métaphysique* : c'est manquer de *paideia* (apaideusia) que « de ne pas discerner entre les choses qu'on doit chercher à démontrer, et celles qu'on ne doit pas démontrer du tout[895] ». Certaines traductions font de ce « manque de *paideia* » un « manque de lumière », ce qui n'est pas sans rappeler les comparaisons classiques entre la lumière et l'intuition. Ici, du fait du lien étymologique entre la *paideia* et l'éducation, le terme désignerait une intuition spécifique, le sentiment d'une conformité à un certain idéal de l'humain. Cet idéal de l'humain, nous pouvons sentir lorsqu'il est menacé ou respecté, mais sans parvenir à le définir.

Le projet d'intelligence artificielle, en se donnant pour tâche de définir ce qu'est la raison et donc, selon Aristote, la différence spécifique qui caractérise l'humain, aurait pour ambition de définir cet idéal et de constituer une sorte de miroir de l'humain. Mais, du fait de la « flexibilité interprétative » des résultats d'intelligence artificielle, aucun programme n'emporte l'unanimité quant au fait qu'il représente effectivement une de nos facultés. Au lieu

894 *Ibid.*

895 Aristote, *Métaphysique,* Livre IV, Chapitre 4, traduction Jules Barthelemy-Saint-Hilaire, Paris, Germer-Baillères, 1879.

de se reconnaître dans les machines qui lui sont présentées, l'humain éprouve une inquiétude, analogue à celle que provoquent les fous, que Clément Rosset impute au fait qu'il est mis face à un vide : il n'existe pas d'essence de l'humain, pas plus que des autres existants. Le dément, comme l'automate, offrent deux manières d'être dépourvu de raison, entre lesquelles on cherche une définition de ce que peut signifier être pourvu de raison, sans jamais la trouver. L'effort est vain, car il n'existe pas de « nature » de la raison, pas plus qu'il n'existe de « nature » de l'humain. Ce sont des formes transitoires nées du hasard, qui prennent des traits différents et singuliers à chaque époque, sans que cette évolution ne suive de principes. Un des traits singuliers de notre époque est justement la quête d'une « nature », d'une forme définitive de l'humain. L'humain moderne présuppose que sa définition existe et la cherche en allant remuer certaines zones frontières où il devient difficile de démêler l'humain de l'inhumain : chez les fous et les machines – deux zones qui se rejoignent aujourd'hui, puisque le projet d'intelligence artificielle est aussi l'occasion d'une ambition de réduire la folie, soit par l'explicitation des mécanismes secrets du cerveau, soit en neutralisant les déviations de comportements.

L'« analogie de l'oignon », que Turing présente dans l'article de 1950, illustre bien cette quête, dans des termes qui rappellent la découverte du *cogito* par Descartes. Pour Turing, élucider le fonctionnement d'une faculté revient à la mettre en doute : si nous pouvons l'effectuer sans y penser, ou si une machine peut le faire, c'est que ladite faculté ne contient pas d'esprit. Tout comme Descartes, à force de douter, finit par buter sur le *cogito,* Turing se demande si l'élucidation de toutes les facultés mécanisables permettra de faire surgir l'« esprit réel ». L'« esprit réel » serait ce qui reste une fois que toutes les facultés ont été élucidées, tout comme le *cogito* est ce qui reste une fois le doute porté à son maximum. Mais le *cogito* n'apparaît qu'à celui qui en fait l'expérience. La « trame démonstrative » ne suffit pas à le montrer, il faut en suivre la « trame ascétique », c'est-à-dire l'énoncer « pour son propre compte[896] », l'effectuer pour soi. Un élève à qui j'enseigne le texte de Descartes peut me persuader qu'il en fait l'expérience, mais je ne peux pas me mettre à sa place pour le vérifier. C'est le problème dit de l'ange de Tobie : « l'observation extérieure ne suffit pas pour accéder à autrui en tant qu'autrui[897] ». Turing semble penser que toutes les facultés de l'esprit pourront être décrites mécaniquement, autrement dit que l'« esprit réel » ne surgira pas. Et en effet, si l'« esprit réel » ne se manifeste que par l'expérience subjective, un spectacteur extérieur ne peut voir que les machines en sont dépourvues. Les zombies sont possibles : nous pouvons imaginer

896 Michel Foucault, « Mon corps, ce papier, ce feu », *op. cit.*, p. 264.
897 Alain de Libera, *L'invention du sujet moderne : cours du Collège de France, 2013-2014, op. cit.* , cours du 19 juin 2014.

des entités qui feraient tout ce que font les humains, de la même manière, sans jamais « y penser », sans conscience, et sans que les humains ne voient de différence.

Comme il est impossible d'objectiver la présence de subjectivité, rien ne fonde le fait que nous attribuons, ou non, de la subjectivité à une entité. Après plusieurs siècles d'un « grand partage » entre un monde de la « culture » réservé aux humains et une « nature » dénuée de subjectivité[898], les frontières se brouillent. Animaux, plantes et fleuves pourraient se voir dotées de subjectivité. Les choses vont encore plus loin avec les *chatbots,* puisqu'il s'agit de matière dite inanimée. Surtout, pour peu qu'ils soient agencés de la bonne manière, n'importe quels éléments du monde peuvent composer un *chatbot* et donc être amenés à passer pour un interlocuteur. Nous explorons plus avant cette question sous la forme d'une fiction, la confession d'un ingénieur chargé de programmer et maintenir un *chatbot* simulant l'écriture de Jacques Derrida, et découvrant *Circonfession,* où Derrida s'imagine répondre à un logiciel qui simulerait sa pensée[899]. En faisant parler un mort, le *chatbot* trouble la distinction entre l'animé et l'inanimé : le vivant et l'humain ont-ils le monopole de la pensée ou bien la matière inerte peut-elle y participer ? Le vivant est-il une condition nécessaire à l'émergence de la pensée, ou bien est-il possible de répéter l'évènement du surgissement de la pensée sans passer par les voies du vivant, en n'utilisant que de la matière inanimée ?

Le projet d'intelligence artificielle se donne pour tâche d'élucider et de reproduire les conditions de la génèse d'une pensée. À ceux qui pourraient objecter que les questions d'origine se situent hors du périmètre de la science, les chercheurs en intelligence artificielle répondent, comme Hilbert l'affirmait au sujet des mathématiques, en réponse à Emil du Bois Reymond, qu'« il n'y a pas d'*ignorabimus,* nous pouvons toujours trouver une réponse à une question pourvu qu'elle ait un sens[900]. » En écrivant que « […] tous les aspects de l'apprentissage ou tout autre caractéristique de l'intelligence peuvent en principe être décrits d'une manière si précise qu'une machine peut être fabriquée pour les simuler[901] », les auteurs de l'appel à projets de Dartmouth affirment à leur tour qu'en sciences cognitives, il n'y a pas *d'ignorabimus, aucun* aspect de l'intelligence ne résiste à la description et il est toujours possible de décrire une caractéristique de l'intelligence d'une manière si précise qu'une machine puisse être fabriquée

898 Philippe Descola, *Par-delà nature et culture,* Paris, Gallimard, 2005.

899 Le texte a été rédigé en 2017. Depuis, des chercheurs ont créé, avec son aide, un *chatbot* simulant les textes de Daniel Dennett. Shayla Love, « In Experiment, AI Successfully Impersonates Famous Philosopher », *Motherboard,* 26 juillet 2022, https://www.vice.com/en/article/epzx3m/in-experiment-ai-successfully-impersonates-famous-philosopher page consultée le 30 juillet 2022.

900 David Hilbert, « Problèmes de fondation des mathématiques », *in* Jean Largeault, *Intuitionnisme et théorie de la démonstration, op. cit.*, p. 185.

901 John McCarthy et al., « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 », *op. cit.*

pour la simuler. Autrement dit, dans le domaine de l'intelligence, il n'y a pas non plus de question sans réponse. Leurs successeurs sont encore plus explicites puisque pour eux, les questions qu'Emil du Bois Reymond plaçait hors du périmètre de la science sont justement celles qui guident leurs recherches : l'origine de la matière, l'origine de la vie, et tout particulièrement l'origine de la conscience. Ce sont, pour Edward Feigenbaum, ces « big magnificent questions[902] » qui font la « destinée manifeste de l'informatique[903] », et pour Yann Le Cun, les rares « questions qui nous intéressent[904] ». Ce faisant, ils se laissent emporter par l'illusion, dénoncée par Kant, que l'expérience pourrait permettre d'élucider les questions d'origine.

C'est en s'interrogeant sur l'origine du monde, et en cherchant à aller « au fond des choses[905] », que Kant bute sur une première antinomie : que le monde ait un commencement ou qu'il n'en ait pas, « la thèse comme l'antithèse peuvent être démontrées par des preuves également lumineuses, claires et irrésistibles » et « la raison se voit ainsi divisée d'avec elle-même[906] ». La lecture de Hume le conduit à déceler une aporie similaire au sujet de la causalité et achève de le « tirer de son sommeil dogmatique[907] ». Dans un cas comme dans l'autre, c'est la notion d'origine ou de création, l'idée qu'entre deux instants successifs, quelque chose *de plus* apparaisse, qui plonge la raison dans la perplexité.

La *Critique de la raison pure* liste les apories auxquelles se heurtent la raison. Trois « Idées » – l'âme, le monde et Dieu – conduisent à quatre antinomies : l'antinomie des limites du monde (dans le temps ou dans l'espace), l'antinomie de la composition (le monde est composé de parties élémentaires indivisibles, ou bien la divisibilité de la matière est infinie), l'antinomie de la liberté (ou bien tout arrive selon les lois de la nature et tout est déterminé ou bien l'âme est à l'origine de ses actions et la liberté existe) et l'antinomie de Dieu (ou bien il existe dans le monde un être absolument nécessaire, qui en est la cause, ou bien aucun être nécessaire ne cause le monde, ni au sein du monde, ni en dehors). Kant réserve un traitement différent aux deux premières antinomies, qu'il juge insolubles (la thèse et l'antithèse se réfutent mutuellement sans que l'une ne l'emporte sur l'autre) et aux deux dernières, où les deux

902 Len Shustek, « An interview with Ed Feigenbaum, *Communications of the ACM, op. cit.*, p. 41-45.
903 Edward Feigenbaum, « Somes Challenges and Grand Challenges for Computational Intelligence », *Journal of the ACM, op. cit.*, p. 39.
904 Yann LeCun, « L'intelligence artificielle dans nos têtes, avec Yann LeCun et Enki Bilal », *Youtube, op. cit.*
905 « Les antinomies n'étaient pas des supercheries, mais elles allaient au fond des choses, sous la supposition que les phénomènes et un monde sensible qui les comprend tous en lui-même seraient des choses en soi. » Emmanuel Kant, *Critique de la raison pure,* B 535, cité et traduit par Paul Clavier, *op. cit.*, p. 9.
906 Emmanuel Kant *Prolégomènes à toute métaphysique future,* § 52a, cité et traduit par Paul Clavier, *op. cit.*, p. 88.
907 Emmanuel Kant *Prolégomènes à toute métaphysique future,* § 50, cité et traduit par Paul Clavier, *op. cit*.

propositions sont vraies, à condition qu'on les applique à des objets différents (choses en soi ou phénomènes). Dans l'ensemble des cas, la même erreur est à l'œuvre : les Idées (l'âme, le monde, Dieu) sont considérées comme des phénomènes – et donc conditionnées –, alors qu'elles sont inconditionnées. La raison s'égare en cherchant l'inconditionné dans les phénomènes. Comme chaque condition peut être donnée, on en déduit que toutes les conditions sont données, et ensuite que la totalité des conditions est aussi un objet qui se donne, sur lequel on peut raisonner. En ce qui concerne le projet d'intelligence artificielle, cela revient à penser que puisque *chaque* aspect de l'intelligence peut être étudié, alors l'intelligence *en tant que totalité* peut également être étudiée. L'intelligence comme totalité, ou comme principe unificateur, pourrait être étudiée de la même manière que ses aspects particuliers. Cela conduit à des apories qui résonnent avec celles listées par Kant : aporie de l'origine (l'intelligence a un commencement, l'intelligence n'a pas de commencement), aporie de la composition ou modularité (l'intelligence est un agrégat de fonctions indépendantes, l'intelligence est une entité « générale » indivisible), et aporie de la liberté (l'intelligence est une suite d'instructions, l'intelligence est libre et à l'origine de ses productions). Mais le pouvoir de séduction des « Idées » contraint la raison à croire qu'elle peut sortir de ces impasses. Kant lui-même, malgré le travail de la *Critique,* ne semble pas être parvenu à renoncer à trouver des réponses empiriques aux questions de l'origine et de l'unité du monde. Quant aux chercheurs en intelligence artificielle, ils ne sont aucunement impressionnés par ces les apories de l'origine, de la divisibilité et de la liberté. Ils n'y voient que des « défis » à résoudre.

Les chercheurs en intelligence artificielle n'utilisent jamais le terme d'âme. Leur objet est l'esprit (« mind »), la conscience ou l'intelligence. Mais lorsque ces notions désignent une substance sous-jacente expliquant la diversité des phénomènes subjectifs, il s'agit bien de ce que Kant appelle l'« Idée d'âme », l'idée d'une unité absolue du sujet pensant. Au lieu d'en rester à une notion de sujet présupposée par l'unité de mes perceptions (sujet empirique), ils postulent l'existence d'une âme comme substance sous-jacente à toutes les expériences subjectives (sujet transcendant). C'est de cette manière que les transhumanistes interprètent les progrès de l'intelligence artificielle : percer les secrets de l'esprit, c'est permettre l'immortalité par « téléchargement ». Dès la constitution d'une théorie de l'information « libre » (non liée à son support) interprétée comme un « code » plutôt que comme un « signal[908] », et la participation du logiciel à cette prise d'indépendance par rapport au matériel (« principe d'implémentation multiple[909] »), l'informatique a été envisagée selon le dualisme corps / esprit.

908 Mathieu Triclot, *Le moment cybernétique : La constitution de la notion d'information, op. cit.,* p. 26.
909 *Ibid,* p. 149.

Voir le projet d'intelligence artificielle comme l'occasion de sauver nos âmes s'inscrit dans le prolongement naturel de cette tendance historique.

La critique kantienne s'applique aussi à ceux qui pensent, tout à l'opposé, que le projet d'intelligence artificielle permettra de prouver que l'Idée d'âme est une superstition, et que l'esprit n'est qu'une collection de processus sans sujet. En voulant prouver que l'âme n'existe pas, la raison outrepasse autant les limites que lorsqu'elle cherche à prouver qu'elle existe. Dans les deux cas, il s'agit de mettre le doigt sur ce qu'est l'âme – quitte à montrer qu'elle n'est rien. L'Idée d'âme ne peut donner lieu qu'à des spéculations que l'expérience ne peut vérifier – elle est hors du périmètre de la science.

Quittant le périmètre de la science, nous entrons à notre tour dans le domaine de la spéculation. Avec l'intuition, nous faisons l'expérience, selon les mots de Turing en 1938, de jugements « spontanés[910] », autrement dit se produisant « sans cause apparente[911] ». Lewis Carroll, dans la préface de *Sylvie et Bruno,* abonde en ce sens. Rien ne pourrait rendre compte des fulgurances improbables qui s'emparent de lui. L'expérience est peut-être trompeuse et ces intuitions pourraient avoir des causes cachées – c'est l'hypothèse que font les chercheurs en intelligence artificielle. Nous choisissons de considérer l'hypothèse inverse : que l'intuition soit *sans cause*. Malgré la prégnance du principe de raison, nous ne devrions pas être si réticents à admettre l'existence d'effets sans cause puisqu'en amont de la chaîne des causes et des effets, se trouve un effet sans cause : le fait de l'existence elle-même. Les mécanistes les plus rigoureux échouent à inclure dans la régression des causes le fait qu'il existe quelque chose plutôt que rien. Interrogés, ils répondraient peut-être qu'un seul évènement sans cause (comme le *big bang*) a mis en route un mécanisme rigoureux. Mais il suffit d'une exception pour que le principe n'en soit plus un, et que d'autres exceptions puissent être admises : l'émergence de la vie, celle de la conscience, ou encore les comportements humains. Admettre quelques exceptions à la nécessité c'est peindre un monde où il y aurait des miracles – sans qu'ils soient pour autant attribués à un Dieu. Dès lors qu'un, deux, ou trois événements sont *sans raison,* rien n'assure que *chaque événement* ne participe pas de la même contingence. L'apparition de la matière, de la vie et de la conscience pourraient n'être que les exemples les plus spectaculaires d'un hasard qui qualifie tout ce qui vient à l'existence, et en particulier nos idées. Notre

910 Alan Turing, « Systems of Logic Based on Ordinals », in Jack Copeland (ed.), *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and And Artificial Life plus The Secret of Enigma*, *op. cit.*, p. 192

911 « Spontané, -ée, adj. Qui se fait de soi-même, sans avoir été provoqué, qui se produit sans cause apparente », *Centre national de ressources textuelles et lexicales,* https://www.cnrtl.fr/definition/spontan%C3%A9, page consultée le 20 mars 2022.

expérience individuelle de la contingence des phénomènes de l'esprit est une expérience de la contingence du monde en général. Romain Graziani analyse certains états mentaux qui résistent au conditionnement : inspiration, sommeil, plaisir[912]… Chacun a fait l'expérience de ce paradoxe de l'insomnie, où tous nos efforts pour réunir les causes d'un état ne font que l'éloigner d'autant. Cela vaut également pour la joie, dont Clément Rosset a montré qu'elle est sans raison. Il n'y a aucune raison d'être joyeux, et pourtant la joie est là[913]. D'autres phénomènes peuvent illustrer la contingence des phénomènes de l'esprit, comme les phobies ou l'effet placebo. Mais il ne s'agit pas de présenter l'esprit comme « la folle du logis » qui interviendrait pour perturber le cours nécessaire des choses matérielles. Il n'y a pas « intervention » du hasard (souvent rapporté à l'esprit) dans des phénomènes par ailleurs déterminés (généralement rapportés à la matérialité du corps), mais formation de stabilités transitoires depuis un fond hasardeux – et cela vaut pour tous les corps.

Pour concevoir de telles hypothèses sur le monde, il faut aller à rebours de la prééminence qui a été donnée à la nécessité. Pour Platon, comme pour Aristote, la recherche philosophique doit se tourner vers les vérités éternelles. Si la Terre et le monde sublunaire sont soumis au changement, le monde supralunaire qui l'entoure est immuable. La physique classique introduit du mouvement dans le cosmos, mais pas de contingence : les objets suivent des trajectoires déterminées. L'objet de la recherche reste ce qui demeure, les formes éternelles qui régissent le changement. Au fil des découvertes du dix-neuvième et du vingtième siècles, il apparaît que l'univers a une histoire contingente, confirmant les intuitions de Pierce selon lesquelles les « lois de la nature » sont des produits historiques, le résultat d'une « prise d'habitude[914] », et non des « lois » *a priori.* Malgré la mise à mal du modèle de la physique classique, son fantôme ne cesse d'exercer une fascination sans précédent sur la plupart des disciplines scientifiques : biologie, génétique, chimie, sociologie – et bien sûr intelligence artificielle. Pour Giuseppe Longo et Pierre-Emmanuel Tendero, les différentes disciplines ont tenté, à la suite de la physique classique, de reproduire le succès de l'alphabet – la mise au point d'un ensemble fini de composants de base et de règles de composition à même de représenter exhaustivement la variabilité du réel. Mais une telle ambition se heurte à la singularité des objets étudiés. Avec le virage connexionniste, l'histoire de l'intelligence artificielle a évolué, comme celle de la physique, suivie par la chimie, la biologie, la génétique, vers une plus grande prise

912 Romain Graziani, *L'usage du vide, Essai sur l'intelligence de l'action, de l'Europe à la Chine, op. cit.*
913 Clément Rosset, *La force majeure, op. cit.*
914 Charles Sanders Pierce, *Le raisonnement et la logique des choses, Les conférences de Cambrigdge (1898), op. cit.*, p. 309-310.

en compte de la singularité des objets étudiés – sans pour autant aller jusqu'au bout. La différence entre l'intelligence artificielle et les autres disciplines réside dans le fait que la singularité semble être constitutive de l'intelligence. Le recours à un modèle simplifié permet de restituer un peu du mouvement des planètes ou du fonctionnement de la cellule, mais il passe à côté de l'intelligence. Pour l'intelligence, il semble que la singularité soit première, si on la « néglige », on perd l'originalité et ce qui fait le point de vue, c'est-à-dire ce qui constitue un interlocuteur. L'originalité et le point de vue, qui constituent l'interlocuteur, sont premiers par rapport aux notions communes.

Sans aller jusqu'à de telles considérations, qui remettent en question le projet d'intelligence artificielle, les évolutions récentes des algorithmes auraient pu conduire les chercheurs à revoir l'horizon de leur projet. Alors que les modèles se singularisent, l'ambition des chercheurs demeure celle de la mise au point d'un modèle universel. Ils ne cessent pas d'être hantés par le fantôme de la physique classique. Les succès retentissants de leurs dernières inventions les conduisent à renouveler la promesse inaugurée dans les années cinquante : d'ici à peu près vingt ans, les recherches aboutiront à la mise au point de machines effectivement intelligentes. Dans la mesure où cette promesse a été réitérée à l'identique depuis soixante-dix ans, il est tentant d'en déduire que les machines intelligentes ont été, sont, et seront *toujours* pour « dans vingt ans ». Elles sont toujours « pour demain », aussi faut-il prendre différement le sens de ce « pour demain » : elles sont un mythe à la faveur duquel nous élaborons des récits sur l'avenir. Elles servent de catalyseur pour fabriquer « demain » au sens où « demain » désigne l'image de l'avenir.

L'intelligence artificielle comme mythe

Le projet d'intelligence artificielle s'aventure hors des limites de la raison. Il est l'occasion d'un discours sur ce qui est « sans pourquoi ». Il s'agit d'un mythe, dans la mesure où, si l'on s'en tient à l'hypothèse de Deuschle, « le mythe sert à exposer les opinions concernant le devenir[915] ». Socrate, écrit Derrida, « envoye promener les mythes[916] » pour ouvrir le dialogue philosophique, mais il les rappelle à l'occasion[917], lorsque le *logos* est dépassé, devant les questions de genèse. Il faut des mythes pour évoquer l'origine de l'humanité (Atlantide dans le *Timée*), l'origine de l'amour (*Le Banquet*), l'origine de l'écriture (Theut dans *Phèdre*), l'origine de la technique (Promethée)[918]… Aussi la volonté d'apporter une réponse scientifique à ces questions peut-elle être considérée comme une forme d'*hubris,* un acharnement à exercer le *logos* là où il est impuissant.

Du point de vue des chercheurs, la fabrication de machines intelligentes est une entreprise de démystification. Elle permettra de faire tomber les derniers bastions de la superstition que sont les mythes de l'esprit immatériel et de l'intuition non mécanique. En réalité, le projet d'intelligence artificielle a un effet inverse. Tout comme Turing, dans l'article de 1950, faisait preuve d'une créativité tous azimuts pour convaincre son lecteur de l'inexistence de la créativité n'existe pas, le projet d'intelligence artificielle, sous couvert d'une entreprise de démystification, est l'occasion d'un renouvellement de mythes anciens, en particulier le mythe de l'âge d'or. Alors que les Grecs situaient l'âge d'or dans le passé, le projet d'intelligence artificielle le renvoie toujours à l'avenir. Dans les deux cas, il s'agit d'un ailleurs temporel où le pénible renouvellement des conditions matérielles de notre existence serait pris en charge – par la nature ou les machines.

Comparant la Genèse aux mythes Iatmul de Nouvelle-Guinée, Bateson remarque que les mythes répondent moins au problème de l'origine de la matière qu'à celui de *l'origine de*

915 Emile Brehier, « Perceval Frutiger. Les mythes de Platon [compte-rendu] », *Revue des Études Grecques,* 44-207, 1931, p. 348-350.

916 Jacques Derrida, *La pharmacie de Platon,* revue Tel Quel, Éditions du Seuil, 1968, p. 7.

917 « Envoyer promener les mythes, les saluer, les mettre en vacances, leur donner congé, cette belle résolution du *chairein,* qui veut dire tout cela à la fois, sera interrompue deux fois pour acueillir ces 'deux mythes platoniciens', donc, 'rigoureusement originaux'. » *Ibid,* p. 9.

918 Page 233, Derrida évoque la synthèse effectuée par Perceval Frutiger, *Les Mythes de Platon*, Paris, Alcan, 1930. Pour Frutiger, tous les mythes ne traitent pas des questions de genèse. Il y a aussi les mythes allégoriques, les mythes parascientifiques, etc.

l'ordre qui organise cette dernière. Il s'agit de raconter la création des distinctions fondamentales entre la lumière et les ténèbres, l'eau et la terre[919]… Une des composantes essentielles du mythe de l'intelligence artificielle est la notion d'intelligence qui, selon les mœurs de notre époque, définit la hiérarchie des êtres. Selon la « Constitution moderne[920] », les entités considérées comme intelligentes mettent à leur service celles dont l'intelligence est considérée comme moindre (femmes, colonisés, animaux…) ou nulle (petits animaux, plantes). Les premières donnent leur force de travail tandis que les secondes donnent leur chair. L'intelligence ordonne le monde en définissant quelles entités doivent sacrifier leur temps, leur force, voire leur chair, au service de quelles autres entités. Une limite a été posée, hier par la notion d'âme, aujourd'hui avec celle de conscience : si une entité est considérée comme consciente, si elle peut faire l'expérience de la souffrance, alors elle ne sera pas traitée seulement comme un moyen, mais aussi comme une fin en soi. C'est l'enjeu de la controverse de Valladolid : examiner l'âme des Indiens pour statuer sur leur place dans la hiérarchie des êtres et déterminer s'ils peuvent être mis en esclavage[921]. Aujourd'hui, la constitution craque de partout : les femmes ne sauraient plus offrir leur temps et leurs efforts à faciliter la vie des hommes, ni être considérées comme inférieurement intelligentes. La notion de conscience s'étend aux animaux, voire aux plantes, ce qui remet en question l'exploitation à grande échelle de leurs corps. Partout, les ressources viennent à manquer. Il n'y a pas assez d'entités à sacrifier pour que les humains aient le train de vie promis par le « progrès » : énergie, eau et nourriture en abondance, temps libre, santé, moyens de transport, de chauffage, de divertissement, etc. Les ressources manquent et elles « réagissent » à la surexploitation : l'eau, la terre, l'air s'empoisonnent, les animaux disparaissent ou deviennent vecteurs de maladies. Le réchauffement induit par nos activités se traduit par une série croissante de catastrophes : canicules, incendies, inondations, fonte des glaces, tempêtes, inondations… Le progrès promis prend de plus en plus sûrement l'aspect d'un suicide collectif où chacun, accroché à son confort, n'en est plus réduit qu'à espérer que les autres souffriront avant lui. La promesse du progrès ne peut plus être tenue, mais personne, ni ceux qui en jouissent, ni ceux qui luttent pour y accéder, ne semblent vouloir y renoncer. Quelles entités de ce monde exsangue pourraient prendre le relai, faire le sacrifice de leur temps et de leurs corps pour permettre aux masses futures de jouir du progrès ?

919 Gregory Bateson, *op. cit.*, p. 23.
920 Nous empruntons le terme à Bruno Latour, *Nous n'avons jamais été modernes, op. cit.,* p. 63.
921 Jean-Claude Carrière, *La controverse de Valladolid*, Paris, Pocket, 2012.

Devant cette tension, le projet d'intelligence artificielle formule deux promesses. La première, évidente, est que les machines intelligentes résoudront cette équation impossible. Elles relanceront la croissance et le progrès tout en fournissant les solutions techniques qui remédieront à l'épuisement des ressources et à l'empoisonnement du monde. La deuxième, plus subtile, est de donner une voix à la matière dite inanimée. Pris au pied de la lettre, le projet de fabrication de machines intelligentes revient à assembler des métaux, du plastique, de l'électricité, de façon à ce qu'il passe pour un interlocuteur convaincant, autrement dit à agencer de la matière non-humaine de manière à ce qu'elle soit en capacité de *nous répondre* – d'animer de la matière dite inanimée. Cela bouleverserait la « Constitution moderne » en effaçant la frontière qui sépare les entités conscientes des entités inanimées. Étant entendu que nous devons certains égards aux premières tandis que nous pouvons disposer à notre guise des secondes, animer la matière en fabriquant des robots qui pensent viendrait bouleverser notre « constitution » et remettre en question ce que nous nous autorisons à faire avec la matière inanimée. Les chercheurs en intelligence artificielle ne semblent pas percevoir cet enjeu. Pourtant, les auteurs de fiction n'ont cessé de le mettre en scène. Il serait trop long de citer toutes les œuvres qui, à l'instar de *Westworld*[922] ou *Real Humans*[923], figurent des robots qui pensent et amènent ainsi le spectateur à considérer qu'il n'est plus possible d'en faire ce que l'on veut. La vision de leurs souffrances, de leurs amours et de leurs ratiocinations sur la conscience amène le public, selon la vulgate kantienne, à ne plus seulement les considérer comme des moyens, mais aussi comme des fins en soi.

En faisant traverser aux robots la frontière partageant les entités inanimées des entités conscientes, c'est toute la constitution qui est remise en cause. Si je laisse entrer un ou deux robots, est-ce qu'il ne faudra pas admettre l'ensemble des non-humains ? N'importe quel objet inanimé pourra-t-il devenir conscient et bénéficier des mêmes égards ? Est-ce à dire que nous devrions des égards à tout ce qui existe ? Si elle cède pour quelques-uns, la légitimité de la frontière ne tient plus pour personne. Dans *Real Humans*, les anti-robots ressemblent à s'y méprendre aux anti-immigration, et pour cause : en traversant la frontière qui partage les entités conscientes des entités inanimées, en obtenant les mêmes droits que les vivants, les robots viennent non seulement les priver de leur travail, mais surtout les priver de leurs colonies. Voilà qu'il n'est plus possible de les exploiter. Voilà qu'il faudrait avoir des égards vis-à-vis de l'ancien esclave : la matière inanimée, la « nature » dont nous nous sommes crus « maîtres et possesseurs ».

922 Jonathan Nolan, Lisa Joy, *Westworld*, New York, HBO, 2016-2022.
923 Harald Hamrell, Lars Lundström, *Real Humans : 100 % humain,* Stockolm, SVT1, Matador Film, 2012-2014.

La promesse de donner une voix à la matière inanimée ne semble pas vouée à aboutir. Nous avons vu comment l'histoire de l'intelligence artificielle n'offrait qu'une suite de déceptions. À chaque fois qu'une machine a réussi une des tâches qui devait confirmer son intelligence, ce n'est jamais la machine qui a été qualifiée d'intelligente mais la tâche qui a été disqualifiée en tant que test de l'intelligence. Trouver un algorithme capable d'effectuer une tâche cognitive, c'est montrer que cette tâche n'a pas besoin d'esprit pour être effectuée. La quête de l'intelligence peut durer longtemps, puisqu'elle est organisée de manière à ne jamais aboutir. Le but du projet d'intelligence artificielle, ainsi que le défend Marcello Vitali-Rosati, est peut-être d'échouer, puisqu'en échouant il permet de réaffirmer l'existence d'une différence spécifique entre les humains et les non-humains – ou au moins entre les entités intelligentes et les entités non intelligentes. En affirmant l'imminence – pour « demain » – d'un accès des non-humains à l'intelligence, il permet de prendre en compte le fait que la frontière entre humains et non-humains ne tient plus, ou n'a jamais tenu, tout en renvoyant à « demain » l'exigence d'en tirer les conséquences, c'est-à-dire de revoir notre relation avec les choses dites inanimées, et en particulier leur exploitation sans limites.

BIBLIOGRAPHIE

Histoire de l'IA et de l'informatique

BRADLEY, Margaret, « Gaspard-Clair-François-Marie Riche de Prony (1755-1839), Constructeur de ponts », *Bulletin de la SABIX*, « Regards sur des carrières de polytechniciens au XIXe siècle », 48 | 2011.

CREVIER, Daniel, *À la recherche de l'intelligence artificielle,* Paris, Flammarion, 1997 (traduit de l'anglais par Nathalie BUCSEK, *AI, The Tumultuous History of the Search for Artificial Intelligence,* New-York, Basic Books, 1993).

DAVIS, Martin, *The Universal Computer, The Road from Leibniz to Turing,* New-York, W. W. Norton & Company, 2000.

DELUZ, Vincent, « De la clepsydre animée à l'horloge mécanique à automates, entre Antiquité et Moyen Âge », *in* Sophie Madeleine, Philippe Fleury et Karim Saamour (dir.) *Autour des machines de Vitruve. L'ingénierie romaine : textes, archéologie et restitution,* Caen, Presses Universitaires de Caen, 2017, p. 173-194.

GOLDSTINE, Hermann, *The Computer from Pascal to von Neumann*, Princeton, Princeton University Press, 1972.

HODGES, Andrew, *Alan Turing, The Enigma*, Princeton, Princeton University Press, 2014.

— *Alan Turing ou l'énigme de l'intelligence*, Paris, Payot, 2004.

HOLLINGS et al., « The Lovelace-De Morgan mathematical correspondence: A critical re-appraisal », *Historia Mathematica*, vol. 44, Issue 3, Août 2017, p. 202-231.

MARCINKOWSKI, Alexandre et Jerôme WILGAUX, « Automates et créatures artificielles d'Héphaïstos : entre science et fiction », *Techniques et culture*, n°2, 2004, p. 43-44.

MARKOFF, John, *Machines of Loving Grace, The Quest for Common Ground Between Humans and Robots*, New-York, Ecco, 2016.

MAYOR, Adrienne, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology,* Princeton, Princeton University Press, 2018.

MCCORDUCK, Pamela, *Machines Who Think, A Personal Inquiry into the History and Prospects of Artificial Intelligence*, Natick, AK Peters, 2004.

METZ, Cade, *Genius Makers, The Mavericks Who Brought AI to Google, Facebook, and the World*, New-York, Penguin Random House, 2021.

NILSSON, Nils, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambrige MA, Cambridge University Press, 2009.

OLAZARAN, Mikel, « A Sociological Study of the Official History of the Perceptrons Controversy », *Social Studies of Science*, vol. 26, n°3, Août 1996, p. 611-659.

PICCININI, Gualtiero, « The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts's 'Logical Calculus of Ideas Imminent in Nervous Activity », *Synthese*, Berlin, Springer, 141, 2004, p. 175–215.

PICKERING, Andrew, *The Cybernetic Brain. Sketches of another Future*, Chicago, The Chicago University Press, 2010.

PRATT, Vernon, *Thinking Machines: The Evolution of Artificial Intelligence*, Oxford, B. Blackwell, 1987.

— *Machines à penser : une histoire de l'intelligence artificielle* traduit de l'américain par Christian PUECH, Paris, Presses Universitaires de France, 1995.

ROLAND, Alex et Philip SHIMAN, *Strategic Computing, DARPA and the Quest for Machine Intelligence, 1983-1993*, Cambridge MA, The MIT Press, 2002.

SEGAL, Jerôme, *Le zero et le un, histoire de la notion d'information au XX è siècle*, Paris, Éditions Matériologiques, 2011.

SOLOMONOFF, Grace, « Ray Solomonoff and the Dartmouth Summer Research Project in Artificial Intelligence, 1956 », http://raysolomonoff.com/dartmouth/dartray.pdf, page consultée le 20 mars 2021.

TARA, Abraham, « (Physio)logical circuits: The intellectual origins of the McCulloch-Pitts neural networks », *Journal of the History of the Behavioral Sciences,* 38(1), Février 2002, p. 3-25.

TEUSCHER, Christof, *Turing's Connectionism: An Investigation of Neural Network Architectures* Berlin, Springer Science & Business Media, 2012.

TOURNÈS, Dominique, « Perspectives historiques sur les abaques et bouliers », *MathémaTICE*, 51, 2016.

TRICLOT, Mathieu, *Le moment cybernétique : La constitution de la notion d'information,* Seyssel, Éditions Champ Vallon, 2008.

TRUITT, Elly Rachel, *Medieval Robots. Mechanism, Magic, Nature, and Art,* Philadelphie, University of Pennsylvania Press, 2015.

TURNER, Fred, *From Counterculture to Cyberculture*, Chicago, University of Chicago Press, 2006.

Articles et textes fondateurs du projet d'IA (1842-2017)

ANDERSON, James, Edward ROSENFELD, *Neurocomputing: Directions for Research*, Cambridge MA, MIT Press, 1988.

BOOLE, George, *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities* (1854), Cambridge, Cambridge University Press, 2009.

BROOKS, Rodney, « Intelligence without Representation », *Artificial Intelligence*, 47(1-3), 1991, p. 139-159.

COPELAND, Jack, (éd.), *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, 2004.

FUKISHIMA, Kunihiko, « Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position », *Biological Cybernetics* 36, 193–202, 1980.

HEBB, Donald O., *The Organization of Behavior: A Neuropsychological Theory*, New York, Wiley, 1949.

HINTON, Geoffrey, David ACKLEY, Terrence SEJNOWSKI, « A learning algorithm for Boltzmann machines », *Cognitive Science*, 9 (1), 1985, p. 147–169.

HINTON, Geoffrey, David RUMELHART, Ronald WILLIAMS, « Learning representations by back-propagating errors », *Nature*, n°323, 1986, p. 533-536.

HINTON, Geoffrey, Yann LECUN, Yoshua BENGIO, « Deep learning », *Nature*, vol. 521, n°7553, 2015.

HOPFIELD, John, « Neural Networks and Physical Systems with Emergent Collective Computational Abilities », *Proceedings of the National Academy of Sciences*, vol. 79, avril 1982, p. 2554-2558.

HUBEL, David, Torsten WIESEL, « Receptive fields, binocular interaction and functional architecture in the cat's visual cortex », *The Journal of Physiology*, 160, 1962.

KAUFMANN, Arnold, « L'imagination artificielle (heuristique automatique) », *Revue française d'informatique et de recherche opérationnelle,* tome 3, n°3, 1969, p. 5-24.

KRIZHEVSKY, Alex, Ilya SUTSKEVER, Geoffrey HINTON, « ImageNet classification with deep convolutional neural networks », *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems* - Volume 1, 2012, p. 1097-1105.

LECUN Yann, Bernard BOSER, John DENKER, Donnie HENDERSON, R. HOWARD, Wayne HUBBARD, Lawrence JACKEL, « Backpropagation Applied to Handwritten Zip Code Recognition », *Neural Computation*, vol. 1, n°4, 1989, p. 541-551.

LOVELACE, Ada A., L.F. MENABREA, « Sketch of the analytical engine invented by Charles Babbage, by LF Menabrea, officer of the military engineers, with notes upon the memoir by the translator », *Taylor's Scientific Memoirs*, 1842, vol. 3, p. 666-731.

MIKOLOV, Tomas, Kai CHEN, Greg CORRADO, Jeffrey DEAN, « Efficient Estimation of Word Representations in Vector Space », arXiv:1301.3781, 2013.

MIKOLOV, Tomas, Kai CHEN, Greg CORRADO, Jeffrey DEAN, Ilya SUTSKEVER, « Distributed Representations of Words and Phrases and their Compositionality », arXiv:1310.4546, 2013.

MCCARTHY, John, Marvin MINSKY, Nathaniel ROCHESTER, Claude SHANNON, « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 », *AI Magazine,* vol. 27, n°4, 2006.

MCCULLOCH, Warren, Walter PITTS, « A Logical Calculus of the Ideas Immanent in NervousActivity », *The Bulletin of Mathematical Biophysics*, vol. 5, n°4, décembre 1943, p. 115-33.

— « How we know universals, the perception of auditory and visual forms », *Bulletin of Mathematical Biophysics*, 9(3), septembre 1947, p. 127-47.

MCCULLOCH, Warren, Walter PITTS, Humberto MATURANA, Jerome LETTVIN, « What the Frog's Eye Tells the Frog's Brain », *Proceedings of the IRE,* vol. 47, Issue 11, Novembre 1959, p. 1940-1951.

MINSKY, Marvin, Seymour PAPERT, *Perceptrons: An Introduction to Computational Geometry*, Cambridge MA, The MIT Press, 1969.

MOORE, Gordon E., « Cramming More Components Onto Integrated Circuits », *Electronics*, n°38, 19 avril 1965, p. 114-117.

NEWELL Allen, Herbert SIMON, John SHAW, « The Logic Theory Machine », *IRE Transactions on Information Theory*, vol. IT-2, n°3, 1956.

NEWELL Allen, Herbert SIMON, « GPS: A Program That Simulates Human Thought », *in* Edward Feigenbaum et Julian Feldman (dir.), *Computers and Thought*, New York, McGraw-Hill, 1963, p. 279-283.

PÉLISSIER, Aline, Alain TÊTE (éd.), *Sciences cognitives, textes fondateurs (1943-1950) : Wiener, Rosenblueth, Bigelow, McCulloch, Pitts, von Neumann, Hebb, Weaver, Shannon, Turing*, Paris, Presses Universitaires de France, 1995.

ROSENBLATT, Frank, « The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain », *Psychological Review*, vol. 65, n°6, 1958.

RUMELHART, David, James MCLELLAND, PDP Research Group, *Parallel Distributed Processing, Volume 1, Explorations in the Microstructure of Cognition: Foundations*, Cambridge MA, MIT Press, 1986.

— *Parallel Distributed Processing, Volume 2, Explorations in the Microstructure of Cognition: Psychological and Biological Models*, Cambridge MA, MIT Press, 1986.

SEJNOWSKI, Terrence, et Charles ROSENBERG, « NET talk: A parallel network that learns to read aloud », *The Johns Hopkins University EE and CS Technical Report*, Janvier 1986.

SHANNON, Claude, « A Symbolic Analysis of Relay and Switching Circuits », *Transactions of the American Institute of Electrical Engineers*, vol. 57, Issue 12, 1938.

— « A Mathematical Theory of Communication », *The Bell System Technical Journal*, vol. 27, 1948, p. 379–423.

SHANNON, Claude et John MCCARTHY (dir.), *Automata Studies*, « Annals of Mathematics Studies », n°34, Princeton, Princeton University Press, 1956.

SILVER, David, Julian SCHRITTWIESER, Karen SIMONYAN, Ioannis ANTONOGLOU, Aja HUANG, Arthur GUEZ, Thomas HUBERT, Lucas BAKER, Matthew LAI, Adrian BOLTON, Yutian CHEN, Timothy LILLICRAP, Hui FAN, Laurent SIFRE, George VAN DEN DRIESSCHE, Thore GRAEPEL et Demis HASSABIS, « Mastering the game of Go without human knowledge », *Nature*, vol. 550, n° 7676, 19 octobre 2017, p. 354–359.

SILVER, David, Thomas HUBERT, Julian SCHRITTWIESER, Ioannis ANTONOGLOU, Matthew LAI, Arthur GUEZ, Marc LANCTOT, Laurent SIFRE, Dharshan KUMARAN, Thore GRAEPEL, Timothy LILLICRAP, Karen SIMONYAN, Demis HASSABIS, « Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm », arXiv:1712.01815, 2017.

TURING, Alan, « On Computable Numbers, With an Application to the Entscheidungsproblem », Proceedings of the London Mathematical Society, Ser. 2, vol. 42, 1937.

— « Proposals for Development in the Mathematics Division of an Automatic Computing Engine (ACE) » (1945), Com Sci 57, National Physical laboratory, Teddington, UK, 1972.

— « Computing Machinery and Intelligence », *Mind*, vol. 59, n°236, Octobre 1950, p. 433-460.

TURING, Alan, et Jean-Yves GIRARD, *La machine de Turing*, Paris, Éditions du Seuil, 1995.

VON NEUMANN, John, « First Draft of a Report on the EDVAC » (1945), *IEEE Annals of the History of Computing*, Volume 15, n°4, 1993, p. 27-43.

— *L'ordinateur et le cerveau,* traduit de l'anglais par Pascal Engel, Paris, Champs Flammarion, 1996.

WEIZENBAUM, Joseph, « ELIZA – a computer program for the study of natural language communication between man and machine », *Communications of the Association for Computing Machinery*, vol. 9, Issue 1, Janvier 1966, p. 36-45.

— *Computer Power and Human Reason*, New-York, W.H. Freeman and Co., 1976.

VASWANI, Ashish, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N. GOMEZ, Lukasz KAISER, Illia POLOSUKHIN, « Attention Is All You Need », arXiv:1706.03762, 6 décembre 2017.

Ouvrages spécialisés (IA, neurosciences, informatique)

ADAMSON, Smith « Machine Learning and Health Care Disparities in Dermatology » *JAMA Dermatology*, 154 (11), 2018, p. 1247–1248.

ANAND, Kanav, Ziqi WANG, Marco LOOG, Jan VAN GEMERT, « Black Magic in Deep Learning: How Human Skill Impacts Network Training », arXiv:2008.05981, 2020.

ARMSTRONG, Stuart, Kaj SOTALA, « How we're predicting AI – or failing to » *in* Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster (dir.), *Beyond AI: Artificial Dreams*, Pilsen, University of West Bohemia, 2012, p. 52–75.

ARMSTRONG, Stuart, Kaj SOTALA, Seán S. Ó HÉIGEARTAIGH, « The errors, insights and lessons of famous AI predictions – and what they mean for the future », *Journal of Experimental & Theoretical Artificial Intelligence,* 26:3, 2014, p. 317-342.

BIRAN, Or, Kathleen MCKEOWN, « Human-centric justification of machine learning predictions », *in* C. Sierra (dir.), *Proceedings of the twenty-sixth international joint conference on artificial intelligence. Main track,* 2017, p. 1461-1467.

BIRAN, Or, Courtenay V. COTTON, « Explanation and Justification in Machine Learning: A Survey », 2017.

BOGROFF, Alexis, Dominique GUÉGAN, « Artificial Intelligence, Data, Ethics: An Holistic Approach for Risks and Regulations », Documents de travail du Centre d'Economie de la Sorbonne 19012, Université Panthéon-Sorbonne (Paris 1), 2019.

BOLUKBASI, Tolga, Kai-Wei CHANG, James ZOU, Venkatesh SALIGRAMA, Adam KALAI, « Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings », arXiv:1607.06520, 2016.

BURRELL, Jenna, « How the machine 'thinks': Understanding opacity in machine learning algorithms », *Big Data & Society*, janvier-juin 2016, p. 1-12.

CASTELVECCHI, Davide, « Can we open the black box of AI ?», *Nature,* 538, 6 octobre 2016, p. 20-23.

CHELLAPILLA, Kumar, Sidd PURI, Patrice SIMARD, « High Performance Convolutional Neural Networks for Document Processing », *Tenth International Workshop on Frontiers in Handwriting Recognition*, Université de Rennes 1, Oct 2006, La Baule (France), Inria-00112631.

CHOLLET, François, « On the Measure of Intelligence », arXiv:1911.01547, 2019.

CHUNG, Sueyeon, Daniel LEE, Haim SOMPOLINSKY, « Classification and Geometry of General Perceptual Manifolds », *Physical Review X* 8 (3), 2017.

CIRESAN, Dan, Ueli MEIER, Jonathan MASCI, Luca Maria GAMBARDELLA, Jürgen SCHMIDHUBER, « Flexible, High Performance Convolutional Neural Networks for Image Classification », *International Joint Conference on Artificial Intelligence IJCAI*, 2011, p. 1237-1242.

COURGEON, Matthieu, *Marc : modèles informatiques des émotions et de leurs expressions faciales pour l'interaction Homme-machine affective temps réel,* thèse de doctorat en Intelligence artificielle, soutenue à l'Université Paris Sud- Paris XI, 2011.

CRAWFORD, Kate, « Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics », *Science, Technology, & Human Values*, 41(1), 2016, p. 77–92.

— *Contre-atlas de l'intelligence artificielle,* traduit de l'anglais (Australie) par Laurent Bury, Paris, Zulma, 2022.

CRAWFORD, Kate, Ryan CALO, « There is a blind spot in AI research », *Nature*, 13 octobre 2016.

CRAWFORD, Kate, Vladan JOLER, « Anatomy of an AI System », 2018, https://anatomyof.ai/, page consultée le 20 novembre 2020.

CRAWFORD, Kate, Meredith WHITTAKER, Madeleine Clare ELISH, Solon BAROCAS, Aaron PLASEK, Kadija FERRYMAN, *The AI Now Report: The Social and Economic Implications of Artificial Intelligence*, Décembre 2018, https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3.pdf page consultée le 20 novembre 2020.

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

DAHL, George, Junshui MA, Robert SHERIDAN, Andy LIAW, Vladimir SVETNIK « Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships », *Journal of Chemical Information and Modeling*, 55 2, 2015, p. 263-74.

DICARLO, James, Davide ZOCCOLAN, Nicole RUST, « How does the brain solve visual object recognition? », *Neuron,* 9 février, 73 (3), 2012 , p. 415-34.

DICARLO, James, David COX, « Untangling invariant object recognition », *Trends in Cognitive Sciences*, 11 2007, p. 333-341.

ELMAN, Jeffrey, « Distributed Representations, Simple Recurrent Networks, And Grammatical Structure », *Machine Learning*, 7, p. 195–225, 1991.

FEI-FEI, Li, Jia DENG, Wei DONG, Richard SOCHER, Li-Jia LI, Kai LI,« ImageNet: A Large-Scale Hierarchical Image Database », 2009 conference on Computer Vision and Pattern Recognition, 2009.

FEIGENBAUM, Edward, « Somes Challenges and Grand Challenges for Computational Intelligence », *Journal of the ACM,* vol. 50, n°1, Janvier 2003.

GATYS, Leon, Alexander ECKER, Matthias BETHGE, « A Neural Algorithm of Artistic Style », arXiv:1508.06576, 2015.

GOEBEL, Randy, Ajay CHANDER, Katharina HOLZINGER, Freddy LECUE, Zeynep AKATA, Simone STUMPF, Peter KIESEBERG, Andreas HOLZINGER, « Explainable AI: The New 42? », *CD-Make,* 2018.

GOODFELLOW, Ian, Jean POUGET-ABADIE, Mehdi MIRZA, Bing XU, David WARDE-FARLEY, Sherjil OZAIR, Aaron COURVILLE, Yoshua BENGIO, « Generative Adversarial Networks », arXiv:1406.2661, 2014.

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

GUEST, Olivia, Bradley C. LOE, « Levels of Representation in a Deep Learning Model of Categorization », bioRxiv 626374, 2019.

HAMILTON, William, Jure LESKOVEC, Dan JURAFSKY, « Diachronic word embeddings reveal statistical laws of semantic change », arXiv:abs/1605.09096, 2016.

HEWITT, John, Christopher MANNING, « A Structural Probe for Finding Syntax in Word Representations », *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019.

HONG, Ha, Daniel YAMINS, Najib MAJAJ, James DICARLO, « Explicit information for category-orthogonal object properties increases along the ventral stream », *Nature Neuroscience*, 19, 2016, p. 613–622.

HUSTON, Matthew, « Artificial intelligence faces reproducibility crisis », *Science*, 16 février 2018, Vol 358, Issue 6377, p. 725-726.

ISOLA, Phillip, Jun-Yan ZHU, Tinghui ZHOU, Alexei A. EFROS, « Image-to-Image Translation with Conditional Adversarial Networks », arXiv:1611.07004, 2017.

JACOBS, Robert A. and Christopher BATES, « Comparing the Visual Representations and Performance of Humans and Deep Neural Networks », *Current Directions in Psychological Science*, 28, 2019, p. 34-39.

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

JORDAN, Michael, Tom MITCHELL, « Machine learning: Trends, perspectives, and prospects », *Science*, vol. 349, n°6245, 2015, p. 255-260.

KRIEGESKORTE, Nikolaus, « Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing », *Annual Review of Vision Science* 1:1, 2015, p. 417-446.

LAKE, Brenden, Tomer ULLMAN, Joshua TENENBAUM, Samuel GERSHMAN, « Building Machines That Learn and Think Like People », arXiv:1604.00289, 2016.

LAPUSCHKIN, Sebastian, Stephan WÄLDCHEN, Alexander BINDER, Grégoire MONTAVON, Wojciech SAMEK, Klaus-Robert MÜLLER, « Unmasking Clever Hans predictors and assessing what machines really learn », *Nature Communications*, 10, 1096, 2019.

LECUN, Yann, Koray KAVUKCUOGLU, Clement FARABET, « Convolutional Networks and Applications in Vision », *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, 253-256, 2010.

LILLICRAP, Thimothy, Adam SANTORO, Luke MARRIS, Colin AKERMAN, Geoffrey HINTON, « Backpropagation and the brain », *Nature Reviews Neuroscience*, 21, 2020, p. 335–346.

MCGRATH, Thomas, Andrei KAPISHNIKOV, Nenad TOMAŠEV, Adam PEARCE, Demis HASSABIS, Been KIM, Ulrich PAQUET, Vladimir KRAMNIK, « Acquisition of Chess Knowledge in AlphaZero », arXiv:2111.09259, 2021.

MENDOZA, Hector, Aaron KLEIN, Matthias FEURER, Jost Tobias SPRINGENBERG, Frank HUTTER, « Towards Automatically-Tuned Neural Networks », *JMLR: Workshop and Conference Proceedings*, 64:58–65, 2016.

MIKOLOV, Tomas, Ilya SUTSKEVER, Kai CHEN, Greg CORRADO, Jeffrey DEAN, Quoc LE, Thomas STROHMANN, « Learning representations of text using neural networks », NIPS Deep learning workshop 2013 slides, NIPS Deep Learning Workshop 2013.

MONTÚFAR, Guido, Razvan PASCANU, Kyunghyun CHO, Yoshua BENGIO, « On the Number of Linear Regions of Deep Neural Networks », arXiv:1402.1869, 2014.

MORDVINTSEV, Alexander, Christopher OLAH, Mike TYKA, « Inceptionism: Going Deeper into Neural Networks », *Google AI Blog*, 17 juin 2015.

MORDVINTSEV, Alexander, Christopher OLAH, Arvind SATYANARAYAN, Ian JOHNSON, Shan CARTER, Ludwig SCHUBERT, Katherine YE, « The Buildings Blocks of Interpretability », *Distill*, 10.23915/distill.00010, 2017.

MORDVINTSEV, Alexander, Christopher OLAH, Ludwig SCHUBERT, « Feature visualization: How neural networks build up their understanding of images », *Distill*, 10.23915/distill.00007, 2017.

PAYEUR, Alexandre, Jordan GUERGUIEV, Friedemann ZENKE, Blake A. RICHARDS, Richard NAUD, « Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits », *Nature Neuroscience*, 24, 2021, p. 1010–1019.

RADFORD, Alec, Luke METZ, Soumith CHINTALA, « Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks », arXiv:1511.06434, 2016.

SCHMIDHUBER, Jurgen, « Deep Learning in Neural Networks: An Overview », *Neural Networks*, volume 61, janvier 2015, p. 85-117.

SHEN, Yujun, Jinjin GU, Xiaou TANG, Bolei ZHOU, « Interpreting the Latent Space of GANs for Semantic Face Editing », arXiv:1907.10786, 2020.

SHUSTEK, Len, « An interview with Ed Feigenbaum, *Communications of the ACM*, June 2010, vol. 53, n°6, p. 41-45.

SU, Hao, Jia DENG, Li FEI FEI, « Crowdsourcing Annotation for Visual Object Detection », Papers from the AAAI Workshop, Toronto, 2012.

VENUGOPALAN, Janani, Li TONG, Hamid Reza HASSANZADEH, May D. WANG, « Multimodal deep learning models for early detection of Alzheimer's disease stage » *Scientific Reports*, 11, 3254, 2021.

YAMINS, Daniel, James J. DICARLO, « Using goal-driven deep learning models to understand sensory cortex », *Nature Neuroscience*, 19, 2016, p. 356-365.

Ouvrages de vulgarisation, manuels techniques

BANDYOPADHYAY, Avimanyu, *Hands On GPU Computing with Python*, Birmingham, Packt Publishing, 2019.

CHANGEUX, Jean-Pierre, *L'homme neuronal,* Paris, Fayard, 1983.

CORNUÉJOLS, Antoine, Laurent MICLET, Vincent BARRA, *Apprentissage artificiel. Concept et algorithmes*, Paris, Eyrolles, 2018.

FLAJOLET, Philippe, Étienne PARIZOT, « Qu'est-ce qu'un algorithme ? », interstices.fr, 2004.

GÉRON, Aurélien, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, Sebastopol, O'Reilly Book, 2017.

GOODFELLOW, Ian, Yoshua BENGIO, Aaron COUVRILLE, *L'apprentissage profond,* traduit par Fabien Navarro, Salima El Kolei, Benjamin Guedj *al,* Paris, Massot Éditions / Quantmetry, 2018.

JULIA, Luc, *L'intelligence artificielle n'existe pas,* Paris, Éditions First, 2019.

KERNIGHAN, Brian et Dennis RITCHIE, *The C Programming Language,* Upper Saddle River, Prentice Hall, 1988.

KUMAR, Manjit, *Le Grand Roman de la physique quantique, Einstein, Bohr… et le débat sur la nature de la réalité*, Paris, Flammarion, 2012.

LECUN, Yann, Stanislas DEHAENE, Jacques GIRARDON, *La plus belle histoire de l'intelligence, Des origines aux neurones artificiels : vers une nouvelle étape de l'évolution,* Paris, Robert Laffont, 2018.

LECUN, Yann, *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond,* Paris, Odile Jacob, 2019.

RASHID, Tariq, *Make Your Own Neural Network : A Gentle Journey Through the Mathematics of Neural Networks, and Making Your Own Using the Python Computer Language,* CreateSpace Independent Publishing Platform, 2016.

RUSSELL, Stuart et Peter NORVIG, *Artificial Intelligence, A Modern Approach*, Upper Saddle River, Prentice Hall, 2010.

Ouvrages de philosophie et sciences humaines sur l'IA, la technique, l'informatique, les sciences cognitives

ANDERSON, Alan Ross, *Pensée et machine,* Seysell, Champ Vallon, 1993.

ANDLER, Daniel, « Connexionnisme et cognition : A la recherche des bonnes questions », *Revue de synthèse*, vol. 111, 1-2, janvier 1990, p. 95–127.

— « From paleo to neo-connectionism », *in* G. Van Der Vijver (dir.), *Perspectives on Cybernetics,* Dordrecht, Kluwer, 1992, p. 125-146.

— « Philosophy of cognitive science », *in* A. Brenner et J. Gayon (dir.), *French Studies in the Philosophy of Science: Contemporary Research in France*, Berlin, Springer, 2009, p. 255-300.

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

— (dir.) *La cognition, Du neurone à la société,* Paris, Gallimard, 2018.

BACHIMONT, Bruno, *Herméneutique matérielle et artefacture : des machines qui pensent aux machines qui donnent à penser ; critique du formalisme en intelligence artificielle*, thèse de doctorat en Sciences appliquées, sous la direction de Jean Petitot, soutenue à l'Ecole Polytechnique en 1996.

— « Nature, Culture et Artefacture : la place de l'intelligence artificielle dans les sciences cognitives », *Intellectica,* 17, 1993, p. 213-238.

— « L'intelligence artificielle comme écriture dynamique : de la raison graphique à la raison computationnelle » in Jean Petitot et Paolo Fabbri (dir.), *Au nom du sens,* Paris, Grasset, 1999, p. 290-319.

— « Signes formels et computation numérique : entre intuition et formalisme », *Critique de la raison computationnelle*, 2004, publié en ligne http://www.utc.fr/~bachimon/Publications_attachments/Bachimont.pdf page consultée le 12 juillée 2019.

BATES, David, *Enlightenment Aberrations : Error and Revolution in France,* Ithaca NY, Cornell University Press, 2002.

— « Creating Insight: Gestalt Theory and the Early Computer » *in* Jessica Riskin (dir.), *Genesis Redux, Essays in the History and Philosophy of Artificial Life*, Chicago, The University of Chicago Press, 2007, p. 237-259.

— « Cartesian Robotics », *Representations*, 124, 2013.

— « Automaticity, Plasticity, and the Deviant Origins of Artificial Intelligence », *in* David Bates et Nima Bassiri (dir.), *Plasticity and Pathology: On the Formation of the Neural Subject*, New York, Fordham University Press, 2015.

— « Insight in the Age of Automation », *in* Joyce Chaplin et Darrin McMahon (dir.), *Genealogies of Genius,* Londres, Palgrave Macmillan, 2016.

— « Automatisation et erreur », *in* Bernard Stiegler (dir.), *La verité du numérique : Recherche et enseignement supérieur à l'ère des technologies numériques,* Paris, FYP Éditions, 2018, p. 29-40.

— « The political theology of entropy: A Katechon for the cybernetic age », *History of the Human Sciences,* vol. 33, Issue 1, 2020, p. 109-107.

BEAUNE, Jean-Claude, *L 'automate et ses mobiles,* Paris, Flammarion, 1980.

BECHTEL, William, Adele ABRAHAMSEN, *Le connexionnisme et l'esprit : introduction au traitement parallèle par réseaux*, traduit par Joëlle PROUST, Paris, Éditions La Découverte, 1993.

BESNIER, Jean-Michel, *Demain les posthumains : le futur a-t-il encore besoin de nous ?* Paris, Hachette littératures, 2009.

BODEN, Margaret A. (dir.), *The Philosophy of Artificial Intelligence,* New-York, Oxford University Press, 1990.

— *The Creative Mind: Myths and Mechanisms*, Londres, Routledge, 2004.

BOSTROM, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2016.

BOURDEAU, Michel, *Pensée symbolique et intuition,* Paris, Presses Universitaires de France, 1999.

BUCKNER, Cameron, « Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks », *Synthese,* 12, p. 1-34, 2018.

— « Deep Learning: A Philosophical Introduction », *Philosophy Compass,* 14, 2019.

— « The Comparative Psychology of Artificial Intelligences », 2019, http://philsci-archive.pitt.edu/16034/1/The%20Comparative%20Psychology%20of%20Artificial%20Intelligence%204%20-%20with%20figures.pdf, page consultée le 20 mars 2020.

— « Understanding adversarial examples requires a theory of artefacts for deep learning », *Nature Machine Intelligence,* vol. 2, décembre 2020, p. 731-736.

— « Black Boxes or Unflattering Mirrors ? Comparative Bias in the Science of Machine Behaviour », *The British Journal for the Philosophy of Sciences,* Avril 2021.

CARDON, Dominique, *À quoi rêvent les algorithmes,* Paris, Seuil, 2015.

— « Intelligence artificielle » in *Culture numérique,* Paris, Presses de Sciences Po, 2019, p. 385-398.

CARDON, Dominique, Antonio CASILLI, *Qu'est-ce que le Digital Labor ?* Bry sur Marne, INA Éditions, 2015.

CARDON, Dominique, Jean-Philippe COINTET, Antoine MAZIÈRES, « La revanche des neurones, L'invention des machines inductives et la controverse de l'intelligence artificielle », *Réseaux,* n°211, 2018, p. 173-220.

CASILLI, Antonio, *En attendant les robots – Enquête sur le travail du clic,* Paris, Éditions du Seuil, 2019.

CASSOU-NOGUÈS, Pierre, « Lacan, Poe et la cybernétique, ou comment le symbole apprend à voler de ses propres ailes », *Savoirs et clinique,* n°16, février 2013, p. 61-70.

— *Hilbert*, Paris, Les Belles Lettres, 2001.

— *Les démons de Gödel : Logique et folie,* Paris, Éditions du Seuil, 2007.

— *Mon zombie et moi*, Paris, Éditions du Seuil, 2010.

— *Lire le cerveau : Neuro/science/fiction*, Paris, Éditions du Seuil, 2012.

— « Signs, figures and time », *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia* 21, n°1, mars 2006, p. 89-104.

— « Deux figures de l'automate spirituel : Leibniz et Turing », *in* L. FÉDI (dir.), *La migration des concepts*, Paris, L'Harmattan, 2002, p. 51-68.

— « 'Vaucanson androïde' : Jean Perdrizet, la cybernétique et le spiritisme », *in* I. MOINDROT, S. SHIN (dir.), *Transhumanités*, Paris, L'Harmattan, 2013, p.135-152.

— « Le temps et la mémoire, l'homme et la machine : autour de la cybernétique », *Intellectica*, 52, 2009/2, p. 141-159.

— *Les rêves cybernétiques de Norbert Wiener,* Paris, Éditions du Seuil, 2014.

— « Gödel et la thèse de Turing », *Revue d'histoire des mathématiques,* 14, 2008, p. 77-111.

CHALMERS, David, « Facing up to the problem of consciousness », *Journal of consciousness studies,* 2(3), 1995, p. 200-219.

CHAZAL, Gérard, *Le miroir automate. Introduction a une philosophie de l'informatique,* Seyssel, Champ Vallon, 1995.

— *Les réseaux du sens, de l'informatique aux neurosciences,* Seyssel, Champ Vallon, 2000.

CITTON, Yves, *Médiarchie*, Paris, Seuil, 2017.

CLARK, Andy, *Microcognition: philosophy, cognitive science, and parallel distributed processing*, Cambridge MA, MIT Press, 1989.

— *Associative Engines: Connectionism, Concepts, and Representational Change*, Cambridge MA, MIT Press, 2003.

CLARK, Andy, et David CHALMERS, « The extended mind », *Analysis*, vol. 58, n°1, Janvier 1998, p. 7-19.

CRAWFORD, Kate, Trevor PAGLEN, « Excavating AI, The Politics of Images in Machine Learning Training Sets », https://excavating.ai/, page consultée le 20 mars 2021.

CREEMERS, Rogier, « China's Social Credit System: An Evolving Practice of Control », *SSRN Electronic Journal*, Janvier 2018.

DAMASIO, Antonio, *L'erreur de Descartes : la raison des émotions,* Paris, Odile Jacob, 1995.

DE BARROS, Manuela, *Magie et technologie*, Paris, UV Éditions, 2017.

DENNETT, Daniel et Douglas HOFSTADTER (dir.), *Vues de l'esprit, fantaisies et réflexions sur l'être et l'âme,* Paris, InterÉditions, 1987.

DREYFUS, Hubert, *Intelligence artificielle : mythes et limites*, Paris, Flammarion, 1984. (traduit de l'anglais par Rose-Marie VASSALLO-VILLANEAU avec le concours de Daniel ANDLER, *What Computers Can't Do, The Limits of Artificial Intelligence* New-York, Harpers & Row, 1979 (1è éd. 1972).

— *What Computers Still Can't Do: A Critique of Artificial Reason,* Cambridge, MIT Press, 1992 (3è édition).

— « Why Heideggerian AI failed and how fixing it would require making it more Heideggerian », *Philosophical Psychology,* vol. 20, n°2, Avril 2007, p. 247-268.

— « A History of First Step Fallacies », in *Minds and Machines,* n°22, 2012, p. 87-99.

DREYFUS, Hubert, Stuart DREYFUS, « What artificial experts can and cannot do », *AI and Society,* 6 (1), 1992, p. 18-26.

ENSMENGER, Nathan, « Is chess the drosophila of artificial intelligence? A social history of an algorithm », *Social Studies of Science,* 42(1), Février 2012, p. 5-30.

EUBANKS, Virginia, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor,* New-York, St Martin's Press, 2018.

FINN, Ed, *What Algorithms Want: Imagination in the Age of Computing,* Cambridge MA, The MIT Press, 2017.

FLORIDI, Luciano, *Philosophy and Computing: An introduction,* London, Routledge, 1999.

— *The 4th Revolution: How the Infosphere is Reshaping Reality,* Oxford, Oxford University Press, 2014.

FODOR, Jerry, *La modularité de l'esprit : essai sur la psychologie des facultés,* Paris, Les Éditions de Minuit, 1986.

FODOR, Jerry, Zenon PYLYSHYN, « Connectionisme and Cognitive Architecture: A Critical Analysis », *Cognition,* 28, 1988, p. 3-71.

FREY, Carl, Michael OSBORNE, « The Future of Employment: How susceptible are jobs to computerisation? », Oxford, Oxford Martin Program on Technology and Employment, 2013.

GANASCIA, Jean-Gabriel, *L'intelligence artificielle*, Paris, Le Cavalier Bleu, 2007.

— *Le mythe de la singularité, Faut-il craindre l'intelligence artificielle ?* Paris, Éditions du Seuil, 2017.

— *L'intelligence artificielle, Vers une domination programmée ?* Paris, Le Cavalier Bleu, 2017.

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

GASTALDI, Juan Luis, « Why Can Computers Understand Natural Language ? The Structuralist Image of Language Behind Word Embeddings », *Philosophy & Technology,* 14 mai 2020.

GASTALDI, Juan Luis, Luc PÉLISSIER, « The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms », *Interdisciplinary Science Reviews,* vol. 46, 2021, p. 569-590.

GILLIE, Donald, *Artificial Intelligence and Scientific Method*, Oxford, Oxford University Press, 1996.

GITELMAN, Lisa (dir.), *Raw data is an oxymoron*, Cambridge, MIT Press, 2013.

GLOBUS, Gordon, « Derrida and connectionism: Differance in neural nets », in *Philosophical Psychology,* 5 (2), 1992, p. 183-197.

GÖDEL, Kurt, Solomon FEFERMAN, John DAWSON *et al* (éd.), *Kurt Gödel, Collected Works, Volume II, Publications 1938-1974,* Oxford, Clarendon Press, 1990.

GOUTEFANGA, Patrick, *Alan Turing : la "pensée" de la machine et l'idée de pratique,* thèse de doctorat en philosophie, sous la direction de Jean-Michel Vienne, soutenue à l'Université de Nantes, 1999.

HAUGELAND, John, *Artificial Intelligence, The Very Idea,* Cambridge MA, The MIT Press, 1985.

HAYLES, Katherine, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, University of Chicago Press, 2008.

HEUDIN, Jean-Claude, *Les créatures artificielles : des automates aux mondes virtuels,* Paris, Odile Jacob, 2008.

HERRENDSCHMIDT, Clarisse, *Les trois écritures. Langue, nombre, code,* Paris, Gallimard, 2007.

HIPOLITO, Inês, « Gödel on the mathematician's mind and Turing Machines », *E-Logos Electronic Journal for Philosophy,* n°22, 2014.

HOFFMAN, Steve, « Thinking science with thinking machines: The multiple realities of basic and applied knowledge in a research border zone », *Social Studies of Science,* 45(2), avril 2015, p. 242-69.

— « Managing Ambiguities at the Edge of Knowledge: Research Strategy and Artificial Intelligence Labs in an Era of Academic Capitalism », *Science, Technology & Human Values* 42 (4), juillet 2017, p. 703-740.

HOFSTADTER, Douglas, *Gödel, Escher, Bach. Les Brins d'une Guirlande Eternelle,* Malakoff, Dunod, 2000.

HORGAN, T., Tienson, J. (dir.), *Connectionism and the Philosophy of Mind,* Berlin, Springer, 1991.

HSU, Feng-hsiung, *Behind deep blue,* Princeton, Princeton University Press, 2002.

HUSTVEDT, Siri, *Les mirages de la certitude, Essai sur la problématique corps / esprit,* Actes Sud, 2018.

HUTCHINS, Edwin, « Material anchors for conceptual blends », *Journal of Pragmatics 37,* (2005), p. 1555-1557.

KARLSSON, Mikael M., « Do We Think With Our Brains ? », *Intellectica*, n°53-54, 2010/1-2, p. 67-94.

KLEIN, Gérard, « Préface », *in* Ian M. BANKS, *Excession,* Paris, Le Livre de poche, 2002.

KOESTLER, Arthur, *The Ghost in the Machine,* Londres, Hutchinson, 1967.

KRAJEWSKI, Stanislaw, « On Gödel Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in Any Attempt to 'Out-Gödel' the Mechanist », *Fundamenta Informatica,* n°81, 2007, p. 1-9.

KURZWEIL, Ray, *The Singularity Is Near: When Humans Transcend Biology*, New-York, Viking Press, 2005.

— *The Age of Spiritual Machines,* New-York, Viking Press, 1999.

LAFONTAINE, Céline, *L'empire cybernétique : des machines à penser à la pensée machine,* Paris, Éditions du Seuil, 2004.

LASSÈGUE, Jean, *L'intelligence artificielle et la question du continu : remarques sur le modèle de Turing,* thèse de doctorat en philosophie sous la direction de Daniel Andler, soutenue à l'Université Nanterre Paris 10, 1994.

— *Turing*, Paris, Les Belles Lettres, 1998.

— « Turing, entre formel et forme ; remarques sur la convergence des perspectives morphologiques », *Intellectica*, n°35, 2002, p.185-198.

— « L'évolution du constructivisme turingien : de la logique à la morphogenèse », *Intellectica,* n°39, 2004/2, p. 107-124.

LAUMOND, Jean-Paul, *La robotique : une récidive d'Héphaïstos : [leçon inaugurale prononcée le jeudi 19 janvier 2012],* Paris, Fayard / Collège de France, 2012.

LÉVY, Pierre, *L'idéographie dynamique, Vers une imagination artificielle,* Paris, La Découverte, 1991.

LONGO, Anna, *Le jeu de l'induction,* Sesto San Giovanni, Éditions Mimésis, 2022.

LONGO, Giuseppe, « The Difference between Clocks and Turing Machines », Conference *Models of Cognition and Complexity Theory*, invited lecture, Rome, November 1994. Proceedings in *La Nuova Critica*, 29 (1), 1995, p. 31-42.

— « Cercles vicieux, mathématiques et formalisations logiques » *Mathématiques et Sciences Humaines*, n°151, 2000.

LONGO, Giuseppe, Pierre-Emmanuel TENDERO, « L'alphabet, la Machine et l'ADN : l'incomplétude causale de la théorie de la programmation en biologie moléculaire ». in Paul-Antoine Miquel, *Biologie du XXIe siècle : évolution des concepts fondateurs*, Louvain la Neuve, DeBoeck, 2008.

LUCAS, John, « Minds, Machines, and Gödel », *Philosophy*, n°36, 1961, p. 112-127.

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

MAITRA, Keya, Jennifer MCWEENY (dir.), *Feminist Philosophy of Mind*, Oxford, Oxford University Press, 2022.

MALABOU, Catherine, *Que faire de notre cerveau ?* Paris, Bayard, coll. « Le temps d'une question », 2004.

— *Ontologie de l'accident, Essai sur la plasticité destructrice*, Paris, Léo Scheer, coll. « Variations », 2009.

— *Métamorphoses de l'intelligence,* Paris, Presses Universitaires de France, 2017.

MALLAT, Stephane, « Understanding deep convolutional networks », *Philosophical Transactions of the Royal Society*, vol. 374, Issue 2065, 13 avril 2016.

MARCUS, Gary, « Deep Learning: A Critical Appraisal », arXiv:1801.00631, 2 janvier 2018.

— *The Algebraic Mind, Integrating Connectionism and Cognitive Science,* Cambridge MA, Bradford Books / MIT Press, 2019.

MARCUS, Gary, Ernest DAVIS, *Rebooting AI, Building Artificial Intelligence We Can Trust*, New York, Vintage, 2019.

MARX, Karl, « Fragment sur les machines », in *Manuscrits de 1857-1858* (« Grundrisse »), Paris, Les Éditions sociales, 2011, p. 660-662.

MAZIÈRES, Antoine, *Cartographie de l'apprentissage artificiel et de ses algorithmes,* thèse soutenue à l'Université Paris Diderot, sous la direction de Jean-Philippe COINTET, 2016.

MINSKY, Marvin, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Simon & Schuster, 2007.

— *The Society of Mind*, Simon & Schuster, 1986.

— « A Framework for Representing Knowledge », *in* Winston (dir.), *The Psychology of Computer Vision*, New York, McGraw-Hill, 1975.

— *Semantic Information Processing*, Cambridge MA, The MIT Press, 1968.

— « Draft of a Proposal to ARPA for Research on Artificial Intelligence at MIT, 1970-1971 », *Artificial Intelligence Lab Publication*, Cambridge, MIT, 1970.

MORAVEC, Hans, *Mind Children: The Future of Robot and Human Intelligence,* Cambridge MA, Harvard University Press, 1988.

— « Rise of the Robots », *Scientific American*, 1er février 2008.

MORI, Masahiro, « La vallée de l'étrange », traduit par Isabel Yaya, *Gradhiva*, vol. 15, 2012b, p. 26-33.

NEGARESTANI, Reza, *Intelligence and Spirit*, Falmouth, Urbanomic, 2018.

OLLION, Étienne, Julien BOELAERT. « Au delà des *big data.* Les sciences sociales et la multiplication des données numériques », *Sociologie*, vol. 6, n°3, 2015, p. 295-310.

O'CONNELL, Mark, *To be a Machine: Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death*, Londres, Granta Books, 2017.

O'NEIL, Cathy, *Algorithmes, la bombe à retardement*, Paris, Les Arènes, 2018.

PEGNY, Maël, Mohamed Issam IBNOUHSEIN, « Quelle transparence pour les algorithmes d'apprentissage machine ? » 2018, hal-01791021.

PENCOLÉ, Marc-Antoine « Nos algorithmes peuvent-ils être plus justes que nous ? », *Revue française d'éthique appliquée*, 2018/1, n°5, p. 67-80.

PENROSE, Roger, *Emperor's New Mind*, Oxford, Oxford University Press, 1989.

PESCHARD, Isabelle, *La réalité sans représentation. La théorie énactive de la cognition et sa légitimité épistémologique*, thèse de doctorat en philosophie des sciences sous la direction de Michel Bitbol, soutenue à l'Ecole Polytechnique en 2004.

PETITOT, Jean, « Hypothèse localiste, modèles morphodynamiques et théories cognitives: remarques sur une note de 1975 », *Semiotica*, vol. 77, 1989, p. 65-119.

PINKAS, Daniel, *La matérialité de l'esprit, la conscience, le langage et la machine dans les théories contemporaines de l'esprit,* Paris, La Découverte, 1995.

PINKER, Steven, Alan PRINCE, « On Language and Connectionism. Analysis of a Parallel Distributed Processing Model of Language Acquisition », *Cognition*, vol. 28, 1988, p. 73-193.

RAMSEY, William, *Representation reconsidered,* Cambridge, Cambridge University Press, 2007.

— « Must Cognition Be Representational ? » In *Synthese*, 194 (11), 2017, p. 4197-4214.

RAMSEY, William, David RUMELHART, Stephen STICH (dir.), *Philosophy and Connexionist Theory,* Routledge, 2016.

ROUVROY, Antoinette, Thomas BERNS, « Le nouveau pouvoir statistique, ou quand le contrôle s'exerce sur un réel normé, docile et sans événement car constitué de corps 'numériques' », *Multitudes,* n°40, 2010/1, p. 88-103.

— « Gouvernementalité algorithmique et perspectives d'émancipation, le disparate comme condition d'individuation par la relation ? », *Réseaux,* n°177, 2013/1, p. 163-196.

ROSSI, Paolo, *Clavis Universalis,* Grenoble, Jérôme Millon, 1993.

— *Les philosophes et les machines, 1400-1700,* Paris, Presses Universitaires de France, 1996.

RYLE, Gilbert, *La notion d'esprit : pour une critique des concepts mentaux,* Paris, Payot, 1978. (Traduit de l'anglais par Suzanne Stern-Gillet, *The Concept of Mind,* Londres, Hutchinson University Library, 1949).

SADIN, Eric, *L'intelligence artificielle ou l'enjeu du siècle : Anatomie d'un antihumanisme radical*, Paris, L'Echappée, 2018.

SCHUBBACH, Arno, « Judging machines: philosophical aspects of deep learning », *Synthese,* 2019, p. 1–21.

SEARLE, John, « Is the Brain a Digital Computer? », *Proceedings and Addresses of the American Philosophical Association,* 64, n°3, 1990, p. 21-37.

— « Minds, brains, and programs », *Behavioral and Brain Sciences,* n°3, septembre 1980, p. 417-24.

— « What Your Computer Can't Know », *The New York Review of Books*, 9 octobre 2014.

SIMONDON, Gilbert, *L'individu et sa genèse physico-biologique*, Grenoble, Jérôme Millon, 1995.

— *Du mode d'existence des objets techniques,* Paris, Aubier, 2012.

SIMONDON, Gilbert, et Jacques GARELLI, *L'individuation à la lumière des notions de forme et d'information,* Grenoble, Jérôme Millon, 2005.

SIMONDON, Gilbert, Nathalie SIMONDON, et Jean-Yves CHATEAU, *Communication et information : Cours et conférences*. Chatou, Éditions de la Transparence, 2010.

SMOLENSKY, Paul, « The proper treatment of connectionism », *The Behavioral and Brain Sciences*, vol. 11, 1988, p. 1-74.

SRNICEK, Nick, *Platform Capitalism*, Cambridge, Polity Press, 2017.

STENGERS, Isabelle, Léon CHERTOK, *Le coeur et la raison. L'hypnose en question, de Lavoisier à Lacan*, Lausanne, Payot, 1989.

STIEGLER, Bernard, *La Technique et le Temps*, Paris, Fayard, 2018.

— *La société automatique 1, L'Avenir du travail,* Paris, Fayard, 2015.

TURKLE, Sherry, « Artificial intelligence and psychoanalysis: a new alliance », *in* Stephen Graubard (dir.), *The artificial intelligence debate: false starts, real foundations*, Cambridge MA, The MIT Press, 1988.

VARELA, Francisco, Humberto MATURANA, *Autopoiesis and Cognition: The Realization of the Living*, Dordrecht, D. Reidel Publishing Company, 1980.

VARELA, Francisco J., Eleanor ROSCH, Evan THOMPSON, *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge MA, MIT Press, 1992.

VARELA, Francisco, Jonathan SHEAR (dir.), *The View From Within: First-Person Approaches to the Study of Consciousness*, Upton Pyne, Imprint Academic, 1999.

VAYRE, Jean-Sébastien, *Des machines à produire des futurs économiques : sociologie des intelligences artificielles marchandes à l'ère du big data*, thèse de doctorat en sociologie, sous la direction de Franck Cochoy, soutenue à l'Université de Toulouse Le Mirail – Toulouse II en 2016.

VIAL, Stéphane, *La structure de la révolution numérique : philosophie de la technologie*, thèse de doctorat en philosophie sous la direction de Maria Michela Manzano soutenue à l'Université Paris 5 en 2012.

VINGE, Vernor, « The Coming Technological Singularity, How to Survive in the Post-Human Era », *in* G.A. Landis (dir.), *Vision-21 : Interdisciplinary Science and Engineering in the Era of Cyberspace*, Washington, NASA Publication, 1993, p. 11-22.

VON FOERSTER (éd.), *Cybernetics, circular causal and feedback mechanisms in biological and social systems, Transactions of the Tenth Conference, April 22, 23 and 24, 1953, Princeton*, New-York, Corlies, Macy & Company, 1955.

WAGNER, Pierre, *La machine en logique*, Paris, Presses Universitaires de France, 1998.

WILLIAMS, Alex, Nick SRNICEK, « #Accelerate Manifesto for an Accelerationnist Politics » (2013), *in* Robin MacKay, Armen Avanessian (dir.), *#Accelerate. The Accelerationnist Reader*, Falmouth, Urbanomic, 2014, p. 349-362.

WINOGRAD, Terry, *Understanding Natural Language*, Edinburgh, Edinburgh University Press, 1972.

WEST, David, Larry TRAVIS, « The Computational Metaphor and Artificial Intelligence: A Reflective Examination of a Theoretical Falsework », *AI Magazine,* 12:64-79, Janvier 1991.

WIENER, Norbert, Arturo ROSENBLUETH, Julian BIGELOW, « Behavior, Purpose and Teleology », *Philosophy of Science*, vol. 10, n°1, 1943, p. 18-24.

WIENER, Norbert, *Cybernetics, or Control and Communication in the Animal and the Machine*, Cambridge MA, MIT Press, 1948.

— *The Human use of Human Beings: Cybernetics and Society,* New-York, Doubleday & Company, 1954.

— *God & Golem, Inc: A comment on certain points where cybernetics impinges on religion,* Londres, Chapman & Hall, 1964.

WIENER, Norbert, Pierre-Yves MISTOULON, et Ronan LE ROUX, *Cybernétique et société : L'usage humain des êtres humains,* Paris, Points, 2014.

WINOGRAD, Terry, Fernando FLORES, *L'intelligence artificielle en question*, Paris, Presses Universitaires de France, 1989.

Ouvrages de philosophie générale et sciences humaines

ABRAM, David, *Comment la terre s'est tue, Pour une écologie des sens*, Paris, Les empêcheurs de penser en rond / Éditions de la Découverte, 2013.

AGAMBEN, Giorgio, *Karman, Court traité sur l'action, la faute et le geste,* Paris, Éditions du Seuil, 2018.

ARISTOTE, *Physique*, traduction Henri Carteron, Paris, Les Belles lettres, 1926.

— *Politique*, traduction Jules Tricot, Paris, Vrin, 2008.

— *Métaphysique,* traduction Alexis Pierron et Charles Zevort, Paris, Ebrard, Joubert, 1840.

— *Métaphysique,* traduction Jules Barthelemy-Saint-Hilaire, Paris, Germer-Baillères, 1879.

— *Métaphysique,* traduction Jules Tricot, Paris, Vrin, 2002.

— *Métaphysique,* traduction Annick Jaulin et Marie-Paule Duminil, Paris, Garnier-Flammarion, 2008.

ATTEN, Mark Van, *On Brouwer,* Wadsworth Philosophers Series, 2004.

AUGUSTIN d'Hippone, *Les Confessions,* Paris, Garnier Flammarion, 1964.

AUROUX, Sylvain, *La révolution technologique de la grammatisation. Introduction à l'histoire des sciences du langage*, Liège, Mardaga, 1994.

BACHELARD, Gaston, *L'engagement rationaliste*, Paris, Presses Universitaires de France, 1972.

— *L'intuition de l'instant,* Paris, Le Livre de Poche, 1994.

BALIBAR, Étienne, « Kant, critique du 'paralogisme' de Descartes. Le 'je pense' (Ich denke) comme sujet et comme substance, *Intellectica,* n°57, 2012/1, p. 21-33.

BARROW, John D., « Varying Constants », *Philosophical Transactions of the Royal Society*, 363, p. 2139-2153.

BARTHES, Roland, *La chambre claire, Note sur la photographie*, Paris, Cahiers du cinéma Gallimard, 1980.

— *Le Neutre : Cours et séminaires au Collège de France (1977-1978),* Paris, Éditions du Seuil, 2002.

— *Comment vivre ensemble : simulations romanesques de quelques espaces quotidiens, notes de cours et de séminaires au Collège de France, 1976-1977,* Paris, Éditions du Seuil, 2002.

BATESON, Gregory, *Vers une écologie de l'esprit 1,* traduit de l'anglais par Ferial Drosso, Laurencine Lot et Eugène Simion, Paris, Éditions du Seuil, 1977.

BERGSON, Henri, *Essai sur les données immédiates de la conscience,* Paris, Presses Universitaires de France, 2013.

— *Matière et mémoire*, Paris, Flammarion, 2012.

— *La Pensée et le Mouvant,* Paris, Flammarion, 2014.

— *L'énergie spirituelle,* Paris, Presses universitaires de France, 2017.

BILLETER, Jean-François, *Héraclite, le sujet*, Paris, Allia, 2022.

BOURDIEU, Pierre, *Esquisse d'une théorie de la pratique précédé de Trois études d'ethnologie kabyle*, Genève, Droz, 1972.

— *Raisons pratiques : sur la théorie de l'action,* Paris, Éditions du Seuil, 1994.

— *Le sens pratique,* Paris, Les Éditions de Minuit, 1980.

BOUQUIAUX, Laurence, « Attention et pensée aveugle chez Leibniz », *Etudes philosophiques,* n°171, 2017/1 , p. 87-102.

BOUVERESSE, Jacques, *Qu'est-ce qu'un sytème philosophique ? Cours 2007 et 2008,* Paris, Collège de France, 2012.

BREHIER, Emile, « Perceval Frutiger. Les mythes de Platon [compte-rendu] », *Revue des Études Grecques*, 44-207, 1931, p. 348-350.

— *La théorie des incorporels dans l'ancien stoïcisme,* Paris, Vrin, 1962.

BRENTANO, Franz, *Psychologie descriptive,* Paris, Gallimard, 2017.

CAILLOIS, Roger, *Les jeux et les hommes. Le masque et le vertige,* Paris, Gallimard, 1967.

— *Le mythe et l'homme,* Paris, Gallimard, 1987.

CANGUILHEM, Georges, *Le normal et le pathologique,* Paris, PUF, 2013.

CHAMAYOU, Grégoire, *Théorie du drone*, Paris, La Fabrique, 2013.

— « Avant-propos sur les sociétés de ciblage, Une brêve histoire des corps schématiques », *Jef Klak*, 21 septembre 2015.

CLAVIER, Paul, *Kant, Les idées cosmologiques,* Paris, Presses Universitaires de France, 1997.

COADOU, François, « 'L'automate spirituel', Contribution à une étude sur la formation du concept de liberté chez Spinoza », *Le Philosophoire*, n°17, 2002/2, p. 187-201.

COULIANO, Ioan Petru, *Eros et magie à la Renaissance : 1484,* Paris, Flammarion, 1984.

DAHAN-GAIDA, Laurence (dir.), *Eurêka ! Récits savants de découvertes et d'inventions,* Paris, Hermann, 2016.

DELEUZE, Gilles, *Logique du sens,* Paris, Éditions de Minuit, 1969.

— « Post Scriptum sur les sociétés de contrôles », *in Pourparlers (1972-1990)*, Paris, Éditions de Minuit, 2003.

— *La Voix de Gilles Deleuze*. Cours sur Spinoza du 02/12/80 www2.univ-paris8.fr/deleuze/article.php3 ?id_article=13, page consultée le 27 août 2019.

DELEUZE, Gilles, Félix GUATTARI, *Qu'est ce que la philosophie,* Paris, Éditions de minuit, 1991.

DE LIBERA, Alain, *L'invention du sujet moderne : cours du Collège de France, 2013-2014,* Paris, Vrin, 2015.

DERRIDA, Jacques, *L'écriture et la différence,* Paris, Seuil, 2014.

— *La pharmacie de Platon*, Paris, Éditions du Seuil, 1968

— *Mal d'archive, une impression freudienne*, Paris, Éditions Galilée, 1995.

— *Foi et savoir* suivi de *Le Siècle et le Pardon,* Paris, Éditions du Seuil, 1991.

— *Circonfession*, Paris, Édition des femmes, 1993.

— *Le toucher, Jean-Luc Nancy,* Paris, Éditions Galilée, 2000.

— *Séminaire La bête et le souverain, Volume I (2001-2002),* Paris, Éditions Galilée, 2008.

— *Séminaire La bête et le souverain, Volume II (2002-2003),* Paris, Éditions Galilée, 2008.

DERRIDA, Jacques, Geoffrey BENNINGTON, *Derrida,* Paris, Éditions du Seuil, 2008 (1è éd. 1991).

DESCARTES, René, *Les passions de l'âme,* Paris, Flammarion, 1998.

— *La Recherche de la Vérité par la lumière naturelle*, Paris, Le Livre de Poche, 2010.

— *Règles pour la direction de l'esprit,* Paris, Le Livre de Poche, 2002.

— *Discours de la méthode*, Paris, Bordas, 1991.

— *Méditations métaphysiques,* Paris, Flammarion, 2009.

— *Correspondance, 1*, Paris, Gallimard, 2013.

DESCOLA, Philippe, *Par-delà nature et culture,* Paris, Gallimard, 2005.

DESROSIÈRES, Alain, *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte, 1993.

DETIENNE, Marcel et Jean-Pierre VERNANT, *Les ruses de l'intelligence, la* mètis *des Grecs*, Paris, Flammarion, 1974.

DUMOUCHET, Paul, Jean-Pierre DUPUY (dir.), *L'Auto-organisation. De la physique au politique*, Colloque de Cerisy (1981), Paris, Seuil, 1994.

DUPUY, Jean-Pierre, *Aux origines des sciences cognitives*, Paris, La Découverte, 2005.

ECO, Umberto, *Sémiotique et philosophie du langage,* Paris, Presses Universitaires de France, 1988.

ELIAS, Norbert, *Qu'est-ce que la sociologie,* Paris, Pocket, 1993.

ESFELD, Michael, *La philosophie de l'esprit : Une introduction aux débats contemporains*, Paris, Armand Colin, 2012.

FAGOT-LARGEAULT, Anne, *Ontologie du devenir, L'évolution, l'univers et le temps,* Paris, Odile Jacob, 2021.

FIRTH, J.R., « A synopsis of linguistic theory 1930–1955 », *Studies in linguistic analysis,* Oxford, Blackwell, 1957.

FINKELSTEIN, Gabriel, *Emil du Bois-Reymond: Neuroscience, Self and Society in Nineteenth-Century Germany*, Cambridge MA, The MIT Press, 2013.

FOUCAULT, Michel, « Introduction », *in* Ludwig Binswanger, *Le rêve et l'existence*, Paris, Desclée de Brouwer, 1954.

— *Histoire de la folie à l'âge classique,* Paris, Gallimard, 2007.

— *Surveiller et punir, Naissance de la prison*, Paris, Gallimard, 1975.

— « La folie, l'absence d'œuvre », *Dits et Ecrits*, tome 1, Paris, Gallimard, 1988.

— « Mon corps, ce papier, ce feu », *Dits et écrits*, tome 2, Paris, Gallimard, 1994.

FREUD, Sigmund, *Le délire et les rêves dans la « Gradiva » de W. Jensen*, Paris, Gallimard, 1986.

FRUTIGER, Perceval, *Les Mythes de Platon*, Paris, Alcan, 1930.

GALILÉE, Christiane CHAUVIRÉ, *L'Essayeur de Galilée*, Paris, Les Belles Lettres, 1979.

GALLAND-SZYMKOWIAK, Mildred, « Le changement de sens du symbole chez Leibniz et Kant », Revue Germanique Internationale, 4 | 2006 : « Esthétiques de l'Auklärung », p. 73-91.

GOODY, Jack, *La raison graphique, La domestication de la pensée sauvage*, Paris, Éditions de Minuit, 1979.

GRAZIANI, Romain, *L'usage du vide, Essai sur l'intelligence de l'action, de l'Europe à la Chine*, Paris, Gallimard, 2019.

HARAWAY, Donna, *The Haraway Reader,* Hove, Psychology Press, 2004.

— « A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century » in *Simians, Cyborgs and Women: The Reinvention of Nature*, New York, Routledge, 1991, p. 149-181.

HAROCHE, Michel-Pierre (dir.), *L'âme et le corps : philosophie et psychiatrie*, Paris, Plon, 1990.

HEINZMANN, Gerhard, « Quelques aspects de l'histoire du concept d'intuition : d'Aristote à Kant » in Pierre Pellerin (dir.), *Philosophie des mathématiques et théorie de la connaissance : l'œuvre de Jules Vuillemin,* Paris, Albert Blanchard, 2005, p. 297-309.

HEIDEGGER, Martin, *Le Principe de raison*, Paris, Gallimard, 1962.

— *Essais et Conférences*, Paris, Gallimard, 1980.

HOBBES, Thomas, *Léviathan*, Paris, Flammarion, 2017.

HUME, David, *L'entendement, Traité de la nature humaine, livre I et Appendice*, traduction de Philippe BARANGER et Philippe SALTEL, Paris, Flammarion, 1995.

HUSSERL, Edmund, *Méditations cartésiennes, Introduction à la phénoménologie*, Paris, Vrin, 1992.

— *Autour des Méditations cartésiennes : (1929 - 1932) ; sur l'intersubjectivité,* Grenoble, Éditions Jérôme Millon, 1998.

— *Recherches logiques : Prolégomènes à la logique pure*, Presses universitaires de France, 1959.

JAEGER, Werner, *Paideia : la formation de l'homme grec,* Paris, Gallimard, 1988.

JOUËT-PASTRÉ, Emmanuelle, « Jeu et éducation dans les *Lois* », *Cahiers du centre Gustave Glotz*, Année 2000, 100, p. 71-84.

KANT, Emmanuel, *Critique de la raison pure*, traduction Renaut, Paris, Flammarion, 2001.

— *Qu'est ce que les Lumières,* Paris, Flammarion, 2020.

LAKOFF, George, Mark JOHNSON, *Metaphors We Live By,* Chicago, University of Chicago Press, 1980.

LAMY, Julien, *Le pluralisme cohérent de la philosophie de Gaston Bachelard*, thèse de doctorat en philosophie, sous la direction Jean-Jacques Wunenburgere, soutenue à l'Université de Lyon 3, 2014.

LAPLACE, Pierre-Simon de, *Essai philosophique sur les probabilités,* Paris, Hachette BNF, 2020.

LARGEAULT, Jean (éd.), *Intuitionnisme et théorie de la démonstration,* Paris, Vrin, 1992.

LARGEAULT, Jean, *Intuition et intuitionnisme*, Paris, Vrin, 1993.

LATOUR, Bruno, *Nous n'avons jamais été modernes,* Paris, Éditions de la Découverte, 1991.

— *Aramis ou l'amour des techniques,* Paris, Éditions de la Découverte, 1992.

— *L'espoir de Pandore, Pour une version réaliste de l'activité scientifique*, traduit de l'anglais par Didier Gille, Paris, Éditions de la Découverte, 2007.

— *Sur le culte moderne des dieux faitiches* suivi de *Iconoclash,* Paris, Les empêcheurs de penser en rond / Éditions de la Découverte, 2009.

— *Cogitamus : Six lettres sur les humanités scientifiques,* Paris, Éditions de la Découverte, 2010.

— *Chroniques d'un amateur de sciences,* Paris, Presses des Mines, 2013.

— *Face à Gaïa, Huit conférences sur le nouveau régime climatique,* Paris, Les Empêcheurs de penser en rond / Éditions de la Découverte, 2015.

LATOUR, Bruno, Pierre LEMONNIER, « Genèse sociale des techniques, genèse technique des humains », *in* Bruno LATOUR et Pierre LEMONNIER (dir.), *De la préhistoire aux missiles balistiques, L'intelligence sociale des techniques,* Paris, Éditions de la Découverte, 1994, p. 11-24.

LEIBNIZ, Gottfried W., « Lettre à Gallois 1677 », *in* C. I. Gerhardt (éd.), *Leibnizen Mathematische Schriften,* Berlin, Verlag von Asher & Comp., 1849.

— « Préface à la science générale », *in* Louis Couturat (éd.), *Opuscules et fragments inédits de Leibniz : extraits des manuscrits de la Bibliothèque royale de Hanovre*, Paris, Félix Alcan, 1903, p. 153-157.

— *Œuvres de G. W. Leibniz* éditées par Lucy Prenant, Paris, Aubier-Montaigne, 1972.

LÉVI-STRAUSS, Claude, « Introduction à l'œuvre de Marcel Mauss » *in* Marcel Mauss, *Sociologie et Anthropologie*, Paris, PUF, 1950.

— *Anthropologie structurale*, Paris, Plon, 1958.

LEROI-GOURHAN, *Le geste et la parole, Tome 1 : Technique et langage*, Paris, Albin Michel, 1975.

— *Le geste et la parole, Tome 2 : La Mémoire et les Rythmes,* Paris, Albin Michel, 1965.

LÉVINAS, Emmanuel, *Théorie de l'intuition dans la phénoménologie de Husserl*, Paris, Vrin, 2001.

LINCOLN, Bruce, *Gods and Demons, Priests and Scholars: Critical Explorations in the History of Religions*, Chicago, The University of Chicago Press, 2012.

LYOTARD, Jean-François, *L'inhumain. Causeries sur le temps,* Paris, Klincksieck, 2014.

MACÉ, Arnaud, « La naissance de la nature en Grèce ancienne », Haber et Macé (dir.), *Anciens et Modernes par-delà nature et société*, Besançon, Presses Universitaires de Franche-Comté, 2012, p. 47-84.

MACHEREY, Pierre, *Querelles cartésiennes*, Villeneuve-d'Ascq, Presses Universitaires du Septentrion, 2014.

MATTEI, Jean-François, *Encyclopédie philosophique universelle – Les œuvres philosophiques, t. 1*, Paris, PUF, 1992.

MEILLASSOUX, Quentin, *L'inexistence divine,* thèse de doctorat en philosophie, sous la direction de Bernard Bourgeois, soutenue à l'Université Paris I, 1996.

— *Après la finitude, essai sur la nécessité de la contingence*, Paris, Éditions du Seuil, 2006.

— « Temps et surgissement ex-nihilo », conférence donnée à l'ENS, avril 2006.

— « Soustraction et contraction, à propos d'une remarque de Deleuze sur Matière et mémoire », *Philosophie*, n°96, 2008/1, p. 67-93.

— *Deuil à venir, Dieu à venir*, Paris, Ismaël, 2017.

NIETZSCHE, Friedrich, *Pour une généalogie de la morale*, in *Œuvres,* Paris, Flammarion, 1996.

— *Par delà bien et mal*, in *Œuvres,* Paris, Flammarion, 2000.

PASCAL, Blaise, *Pensées*, Paris, Bordas, 1966.

PAUGAM, Serge (dir.), *Les 100 mots de la sociologie*, Paris, Presses Universitaires de France, coll. « Que Sais-Je ? », 2018.

PETITMENGIN, Claire, *L'expérience intuitive,* Paris, L'Harmattan, 2001.

PIERCE, Charles Sanders, *Le raisonnement et la logique des choses, Les conférences de Cambrigdge (1898)*, Paris, Le Cerf, 1995.

PLATON, *Ménon,* Paris, Flammarion, 1991.

— *Le banquet,* Paris, Flammarion, 1992.

— *Cratyle,* Paris, Flammarion, 1998.

— *Les Lois, Livres I à VI,* Paris, Flammarion, 2006.

— *Les Lois, Livres VII à XII,* Paris, Flammarion, 2006.

— *Phèdre,* Paris, Flammarion, 2012.

— *La République,* Paris, Flammarion, 2016.

POINCARÉ, Henri, *Science et Méthode*, Paris, Flammarion, 1947.

PRADELLE, Dominique, *Intuition et Idéalités,* Paris, Presses Universitaires de France, 2020.

RABOUIN, David « Penser comme un pied », Intuitive notebook, diagrams, drawing and spaces, *Revue du laboratoire des intuitions*, numéro -1, Annecy, Éditions ESAAA, Mai 2014.

REY, Olivier, *Quand le monde s'est fait nombre,* Paris, Stock, 2017.

— « La confusion des lois » *in* Giuseppe Longo (dir.), *Lois des dieux, des hommes et de la nature. Éléments pour une analyse transversale*, Spartacus, p. 61-73, 2017.

ROUSTANG, François, *Influence,* Paris, Éditions de Minuit, 1991.

— *Jamais contre d'abord, la présence d'un corps,* Paris , Odile Jacob, 2019.

ROSSET, Clément, *Logique du pire,* Paris, Presses Universitaires de France, 1971.

— *Le réel et son double,* Paris, Gallimard, 1976.

— *Le Réel. Traité de l'idiotie,* Paris, Les Éditions de Minuit, 1977.

— *L'objet singulier,* Paris, Les Éditions de Minuit, 1979.

— *La force majeure,* Paris, Les Éditions de Minuit, 1983.

— *Principes de sagesse et de folie,* Paris, Les Éditions de Minuit, 1991/2004.

— *L'école du réel,* Paris, Les Éditions de Minuit, 2008.

RUBIO, Josep, *Raymond Lulle, le langage et la raison : une introduction à la genèse de l'Ars,* Paris, Vrin, 2007.

SLOTERDIJK, Peter, *Règles pour le parc humain* suivi de *La domestication de l'Être,* Paris, Fayard / Mille et une nuits, 2010.

— *Colère et temps,* Paris, Fayard / Pluriel, 2011.

SNELL, Bruno, *La Découverte de l'esprit. La genèse de la pensée européenne chez les Grecs,* Paris, Éditions de l'Éclat, 1994.

SPINOZA, Baruch, Bernard PAUTRAT, *Éthique*, Paris, Seuil, 2014.

SUPIOT, Alain, *La Gouvernance par les nombres, Cours au Collège de France (2012-2014)*, Paris, Fayard, 2019.

VERNANT, Jean-Pierre, « Le Mythe prométhéen chez Hésiode », *Mythe et société en Grèce ancienne*, Paris, Éditions Maspero, 1974, p. 177-194.

— « L'Univers, les Dieux, les Hommes, Récits grecs des origines », *Œuvres*, Tome 1, Paris, Éditions du Seuil, 2007.

— *Pandora, la première femme*, Paris, Éditions Bayard, 2005.

WEBER, Max, *L'éthique protestante et l'esprit du capitalisme*, Paris, Flammarion, 2017.

WHITEHEAD, Alfred North, *An Introduction to Mathematics*, New York, H. Holt, 1911.

Œuvres de fiction, littérature

ASIMOV, Isaac. *I, Robot*. London, HarperVoyager, 2013.

BAUDELAIRE, Charles, *Le Spleen de Paris, Petits poèmes en prose*, Paris, Gallimard, 2006.

BORGES, Jorge, « La loterie à Babylone », in *Fictions*, Paris, Gallimard, 1999.

BUTLER, Samuel, Valery LARBAUD, *Erewhon ou De l'autre côté des montagnes,* Paris, Gallimard, 1981.

CAPEK, Karel, *RUR: Rossum's Universal Robots,* Paris, Éditions de La Différence, 2011.

CARROLL, Lewis, *Through the looking glass*, Londres, Penguin Books, 1994.

— *De l'autre côté du miroir* in *Œuvres*, Paris, Gallimard, 1990.

— *Sylvie and Bruno*, Londres, Macmillan and Co., 1889.

CARRIÈRE, Jean-Claude, *La controverse de Valladolid*, Paris, Pocket, 2012.

DE L'ISLE-ADAM, Auguste de Villiers, Alan RAITT (éd.), *L'Ève future,* Paris, Gallimard, 1993.

GRACQ, Julien, *Préférences*, Paris, José Corti, 1961.

HÉSIODE, *Les travaux et les jours*, Paris, Belles Lettres, 2018.

SIMENON, Georges, *Les vacances de Maigret*, Paris, Le Livre de Poche, 2001.

VONNEGUT, Kurt, *Player Piano: A Novel*, New-York, The Dial Press, 1999.

Films, séries

BROOKER, Charlie, *Black Mirror,* Londres, Channel Four, 2011-2014.

— *Black Mirror,* Los Gatos, Netflix, 2016-2019.

DANIELS, Greg, *Upload,* Seattle, Prime Videos, 2020.

HAMRELL, Harald, Lars LUNDSTRÖM, *Real Humans,* Stockolm, SVT1, Matador Film, 2012-2014.

JOY, Lisa, Jonathan NOLAN, *Westworld,* New York, HBO, 2016-2022.

OSHII, Mamoru, *Ghost in the shell*, Musashino, I.G. Productions, 1995.

PFISTER, Wally, *Transcendance*, Los Angeles, Alcon Entertainment, 2014.

Vidéos, documentaires, conférences

« We talked to Sophia – the first-ever robot citizen that once said it would 'destroy humans' », chaîne *Youtube* de *Tech Insider,* 28 décembre 2017, https://www.youtube.com/watch?v=78-1MlkxyqI, page consultée le 20 juillet 2020.

« Hot Robot At SXSW Says She Wants To Destroy Humans | The Pulse », chaîne *Youtube* de *CNBC,* 16 mars 2016, https://www.youtube.com/watch?v=W0_Dpi0PmF0, page consultée le 20 juillet 2020.

GODARD, Jean-Luc, *Je vous salue Sarajevo,* Toulouse, Éditions ECM Cinema, 1993.

HINTON, Geoffrey, « The Foundations of Deep Learning », *Youtube,* 7 février 2018.

KOHS, Greg, *AlphaGo*, Netflix, 2018, 90 minutes.

LECUN, Yann, « L'apprentissage profond : une révolution en intelligence artificielle », leçon inaugurale au Collège de France, 4 février 2016 https://www.college-dfrance.fr/media/yann-lecun/UPL7915574462521283497_lecun_20160204_college_de_france_lecon_inaugurale.pdf consulté le 23 décembre 2018.

— « The Epistemology of Deep Learning », *YouTube,* 22 février 2019.

— « Les émotions sont inséparables de l'intelligence », *YouTube,* 18 octobre 2019.

— « L'Intelligence artificielle dans nos têtes, avec Yann Lecun et Enki Bilal », *YouTube*, 30 janvier 2018.

LECUN, Yann, Gary MARCUS, « Debate: 'Does AI Need More Innate Machinery?' », *YouTube*, 20 octobre 2017.

LECUN, Yann, Andrew NG, « Heroes of Deep Learning: Andrew Ng interviews Yann LeCun », *YouTube*, 7 avril 2018.

LENA, Pierre, « D'un univers statique à un univers en devenir », Collège de France, 25 janvier 2007, https://www.college-de-france.fr/site/anne-fagot-largeault/course-2007-01-25-10h30.htm page consultée le 28 mai 2020.

MALLAT, Stéphane, « Application des réseaux de neurones profonds », Cours au Collège de France, 30 janvier 2019, https://www.college-de-france.fr/site/stephane-mallat/course-2019-01-30-09h30.htm, page consultée le 12 juin 2022.

NG, Andrew, Geoffrey HINTON, « Heroes of Deep Learning: Andrew Ng interviews Geoffrey Hinton », https://www.youtube.com/watch?v=-eyhCTvrEtE, page consultée le 4 octobre 2019.

VITALI-ROSATI, Marcello, « Qu'est-ce qui échappe à l'intelligence artificielle ? », *YouTube*, 26 octobre 2021, https://www.youtube.com/watch?v=RVFCpT6X1j8, page consultée le 20 mars 2022.

Dictionnaires

BAILLY, Anatole, *Dictionnaire Grec-Français, Le Grand Bailly,* Paris, Hachette, 2000.

GAFFIOT, Félix, *Dictionnaire illustré latin-français,* Paris, Hachette, 1937.

LALANDE, André, *Vocabulaire technique et critique de la philosophie,* Paris, PUF / Quadrige, 2016.

THOMPSON, Della, *The Oxford Dictionary of Current English,* Oxford, Oxford University Press, 1993.

Articles de presse

« China Focus: AI beats human doctors in neuroimaging recognition contest », *Xinhua Net*, 30/06/2018, http://www.xinhuanet.com/english/2018-06/30/c_137292451.htm, page consultée le 1er février 2021.

BOTTICELLI, Manon, « Un tableau produit par intelligence artificielle vendu chez Christie's plus de 40 fois son estimation », *France Télévisions,* 27 octobre 2018, https://www.francetvinfo.fr/culture/arts-expos/un-tableau-produit-par-intelligence-artificielle-vendu-chez-christies-plus-de-40-fois-son-estimation_3368107.html

BRITTAIN, Blake, « U.S. appeals court says artificial intelligence can't be patent inventor », *Reuters,* 5 août 2022, https://www.reuters.com/legal/litigation/us-appeals-court-says-artificial-intelligence-cant-be-patent-inventor-2022-08-05/?utm_source=substack&utm_medium=email page consultée le 25 août 2022.

BRUEL, Benjamin, « Une intelligence artificielle de Facebook a accidentellement inventé son propre langage », *France 24,* 20 juin 2017, https://www.france24.com/fr/20170620-une-intelligence-artificielle-facebook-a-accidentellement-invente-son-propre-langage, page consultée le 2 février 2021.

CAMPBELL, Ian Carlos, « The Apple Card doesn't actually discriminate against women, investigators say », *The Verge,* 23 mars 2021, https://www.theverge.com/2021/3/23/22347127/goldman-sachs-apple-card-no-gender-discrimination page consultée le 20 septembre 2021.

CLARK, Jack, « Artificial intelligence has a sea of dudes problem », *Bloomberg,* 23 juin 2016, https://www.bloomberg.com/news/articles/2016-06-23/artificial-intelligence-has-a-sea-of-dudes-problem

DAS, Sejuti, «Yann LeCun Thrashes GPT-3 — Is The Hype Real?», *Analytics India Magazine,* 28 octobre 2020, https://analyticsindiamag.com/yann-lecun-thrashes-gpt-3-is-the-hype-real/, page consultée le 2 février 2021.

DUFOUR, Audrey, Alice LEDRÉAU, « On est encore loin de l'intelligence humaine ou animale », entretien avec Laurence Devillers et Yann Le Cun, *La Croix,* mardi 2 février 2021, https://www.la-croix.com/Sciences-et-ethique/On-encore-loin-lintelligence-humaine-animale-2021-02-01-1201138254 page consultée le 4 février 2021.

HEAVEN, Will Douglas, « Our weird behavior during the pandemic is messing with AI models », *MIT Technology Review,* 11 mai 2020, https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-

machine-learning-amazon-retail-fraud-humans-in-the-loop/? Page consultée le 20 novembre 2020.

GÉVAUDAN, Camille, « Le jeu de go pour les nuls », *Libération,* 8 mars 2016, https://www.liberation.fr/futurs/2016/03/08/le-jeu-de-go-pour-les-nuls_1438397/ page consultée le 20 novembre 2020.

GOODE, Lauren, « How Google's Eerie Robot Phone Calls Hint at AI's Future », *Wired,* 5 août 2018, https://www.wired.com/story/google-duplex-phone-calls-ai-future/ page consultée le 20 novembre 2020.

LOVE, Shayla, « In Experiment, AI Successfully Impersonates Famous Philosopher », *Motherboard,* 26 juillet 2022, https://www.vice.com/en/article/epzx3m/in-experiment-ai-successfully-impersonates-famous-philosopher page consultée le 30 juillet 2022.

METZ, Cade, « A breakthrough for AI technology: Passing an 8th-grade science test », *Seattle Times,* 4 septembre 2019, https://www.seattletimes.com/business/a-breakthrough-for-ai-technology-passing-an-8th-grade-science-test/, page consultée le 20 novembre 2020.

TAIT, Amelia, « 'I am, in fact, a person': can artificial intelligence ever be sentient? » *The Guardian,* 14 août 2022, https://www.theguardian.com/technology/2022/aug/14/can-artificial-intelligence-ever-be-sentient-googles-new-ai-program-is-raising-questions page consultée le 30 août 2022.

ROOSE, Kevin, « An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. "I won, and I didn't break any rules," the artwork's creator says. », *The New York Times,* 2 septembre 2022, https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html page consultée le 5 septembre 2022.

WILSON, Mark, « AI Is Inventing Languages Humans Can't Understand. Should We Stop It? », *Fast Company,* 14 juillet 2017, https://www.fastcompany.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it page consultée le 2 février 2021.

WIGGERS, Kyle, « Yann LeCun and Yoshua Bengio: Self-supervised learning is the key to human-level intelligence », *VentureBeat,* 2 mai 2020, https://venturebeat.com/2020/05/02/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/ page consultée le 2 mars 2021. L'article est un compte-rendu des interventions de LeCun et Bengio à la conférence ICLR 2020.

Articles de blogs

« Better Language Models and Their Implications », *Open AI*, 14 février 2019, https://openai.com/blog/better-language-models/, page consultée le 1er février 2021.

« Yann LeCun on a vision to make AI systems learn and reason like animals and humans », *Meta AI Blog*, 23 février 2022, https://ai.facebook.com/blog/yann-lecun-advances-in-ai-research/ page consultée le 22 avril 2022.

« Robotique et intelligence artificielle : il faut que les nouvelles technologies bénéficient à tous, selon l'ONU », *ONU Info*, 11 octobre 2017, https://news.un.org/fr/story/2017/10/365972-robotique-et-intelligence-artificielle-il-faut-que-les-nouvelles-technologies, page consultée le 20 juillet 2020.

BARAZIDA, Nir, « Stable Diffusion: Best Open Source Version of DALL·E 2 », *Towards Data Science*, 30 août 2022, https://towardsdatascience.com/stable-diffusion-best-open-source-version-of-dall-e-2-ebcdf1cb64bc page consultée le 3 septembre 2022

BRAHIMI, Idriss, « Un premier retour d'expérience sur AutoML », *Ysance Blog*, 28 septembre 2020, https://blog.ysance.com/un-premier-retour-experience-sur-automl, page consultée le 20 septembre 2021.

BRUNER, Jon, Adit DESHPANDE, « Generative Adversarial Networks for beginners », O'Reilly, 7 juin 2017, https://www.oreilly.com/content/generative-adversarial-networks-for-beginners/, page consultée le 20 septembre 2021.

CHABLANI, Manish, « Word2Vec (skip-gram model): PART 1 - Intuition », *Towards data science*, 14 juin 2017, https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b page consultée le 20 novembre 2020.

CHATONSKY, Grégory, « Capture », http://chatonsky.net/capture/ page consultée le 20 novembre 2020.

GABOURY, Jacob « A Queer History of Computing », *Rhizome*, 9 avril 2013, https://rhizome.org/editorial/2013/apr/9/queer-history-computing-part-three/ page consulté le 20 juillet 2019.

GOODWIN, Ross, « Adventures in Narrated Reality », 19 mars 2016, https://medium.com/artists-and-machine-intelligence/adventures-in-narrated-reality-6516ff395ba3, page consultée le 20 novembre 2020.

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

KNAPTON, Sarah, « AlphaGo Zero: Google DeepMind supercomputer learns 3,000 years of human knowledge in 40 days », *The Telegraph*, 18 octobre 2017, https://www.telegraph.co.uk/science/2017/10/18/alphago-zero-google-deepmind-supercomputer-learns-3000-years/ consulté le 14 septembre 2019.

KUNG-HSIANG, Huang (Steeve), « Word2Vec and FastText Word Embedding with Gensim », *Towards data science*, 4 février 2018, https://towardsdatascience.com/word-embedding-with-word2vec-and-fasttext-a209c1d3e12c, page consultée le 10 août 2021.

KURENKOV, Andrey, « A Brief History of Neural Nets and Deep Learning », http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/ consulté le 20 décembre 2017.

— « LaMDA's Sentience is Nonsense - Here's Why », *Last Week in AI*, 24 juin 2022, https://lastweekin.ai/p/lamdas-sentience-is-nonsense-heres

LEMOINE, Blake, « Is LaMDA Sentient? — an Interview », *Medium*, 11 juin 2022, https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 page consultée le 30 juin 2022.

LEWIS, Mike, Denis YARATS, Devi PARIKH, Dhruv BATRA, «Deal or no deal? Training AI bots to negotiate », *Facebook Engineering*, 14 juin 2017, https://engineering.fb.com/2017/06/14/ml-applications/deal-or-no-deal-training-ai-bots-to-negotiate/, page consultée le 1er février 2021.

MORIN, Violaine « Sophia, robot saoudienne et citoyenne », *Le Monde*, 4 novembre 2017, https://www.lemonde.fr/idees/article/2017/11/04/sophia-robot-saoudienne-et-citoyenne_5210094_3232.html, page consultée le 20 juillet 2020.

REYNOLDS, Anh H. « Convolutional Neural Networks (CNNs) », https://anhreynolds.com/blogs/cnn.html page consultée le 10 novembre 2020.

INDEX NOMINUM

H

I

J

K

L

M

N

O

P

R

S

T

V

W

Z