

IA GÉNÉRATIVES

Guide de survie

Erwan Lamy
Louis-David Benyayer
Isabelle Beyneix
Alara Tascioglu



GUIDE DE SURVIE À L'ÈRE DES IA GÉNÉRATIVES

IA GÉNÉRATIVES

Guide de survie

Erwan Lamy
Louis-David Benyayer
Isabelle Beyneix
Alara Tascioglu



Conception graphique couverture : Éric VIGIER

Conception graphique intérieur et mise en pages : Soft Office

ISBN 9782340-120761

Dépôt légal : octobre 2026

©Ellipses Édition Marketing S.A.

8/10 rue la Quintinie 75015 Paris



Le Code de la propriété intellectuelle et artistique n'autorisant, aux termes des alinéas 2 et 3 de l'article L. 122-5, d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause, est illicite » (alinéa 1^{er} de l'article L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

www.editions-ellipses.fr

SOMMAIRE

- 1. INTRODUCTION ET AVERTISSEMENT 7
 - 1.1. Une technologie impossible..... 10
 - 1.2. Stupéfaction 15
 - 1.3. Au-delà des LLM..... 16
 - 1.4. Le deuil de l’ancien monde..... 19
 - 1.5. Se préparer au nouveau monde..... 22
- 2. L’IA GÉNÉRATIVE, QU’EST-CE QUE C’EST ? 25
 - 2.1. Définir l’IA 27
 - 2.2. Les paradigmes de l’IA..... 32
 - 2.3. Réseaux de neurones 37
 - 2.4. Entraînement 44
 - 2.5. Apprentissage automatique 49
 - 2.6. Deep learning et rétropropagation..... 53
 - 2.7. IA génératives 65
- 3. GRANDS MODÈLES DE LANGAGE 73
 - 3.1. Plongement lexical..... 80
 - 3.2. Attention..... 90
 - 3.3. Classification..... 103
 - 3.4. Ajustements..... 119
 - 3.5. Chain of thoughts..... 126
 - 3.6. Scaling laws..... 130

4. LIEUX COMMUNS	139
4.1. Qualité manquante	145
4.2. Nature déficiente	154
4.3. Simulation	158
4.4. Les LLM racontent n'importe quoi	160
4.5. Des machines logiques et rationnelles	161
4.6. Des bases de données	164
4.7. Des machines sans créativité	165
4.8. De simples outils	167
4.9. Des instruments idéologiques	170
4.10. Des boîtes noires	171
5. LIMITES STRUCTURELLES	175
5.1. Mises au point	177
5.2. Génies stupides	180
5.3. Machines à bullshit	186
5.4. Hallucinations	189
5.5. Indifférence au vrai ?	194
5.6. Source de la normativité	199
5.7. Gentils psychopathes	206
5.8. Quelles solutions ?	213
6. DES BONS USAGES DES IA GÉNÉRATIVES	217
6.1. Coder (Alara Tascioglu, à partir d'un entretien avec Derya Karlı)	220
6.2. Créer (Isabelle Beyneix, entretien avec Guillaume Marien)	231
6.3. Dire le droit (Isabelle Beyneix, entretien avec Pascal Montfort)	248
6.4. Écrire (Erwan Lamy, entretien avec Raphaël Doan)	260
6.5. Enseigner (Louis-David Benyayer)	270
6.6. Entreprendre (Erwan Lamy, entretien avec Younes Boual)	284

Sommaire

6.7.	Plaider (Isabelle Beyneix, avec le concours de Me Julien Guinot-Deléry et de Me Nicolas Laurent).....	298
6.8.	Transformer (Louis-David Benyayer, entretien avec Thibaut Guilluy).....	316
6.9.	Vendre (Alara Tascioglu, à partir d'un entretien avec Rémi Ipekci).....	325
7.	DE LA LECTURE À L'ACTION : VOTRE GRILLE DE DIAGNOSTIC	337
7.1.	Mode d'emploi de la grille.....	339
7.2.	Choisir un mode de délégation.....	342
7.3.	Questionnaire d'orientation.....	345
Index		349
Annexe 1		361
Annexe 2		367
Notes		379

1

INTRODUCTION ET AVERTISSEMENT

Trois ans après son lancement, 900 millions de personnes utilisaient ChatGPT. Aucune autre innovation ne s'est diffusée aussi rapidement dans l'histoire des technologies. Cela n'aurait évidemment pas beaucoup de sens de comparer cela à la vitesse de diffusion de l'électricité, de la voiture ou du téléphone, ou toute technologie qui demande des investissements matériels et financiers pour se déployer. Pour prendre donc des exemples comparables, il a fallu une dizaine d'années à Google et 9 ans à Facebook pour atteindre un nombre d'utilisateurs comparable. Et ChatGPT n'est qu'un chatbot de nouvelle génération parmi d'autres. La famille des GPT d'OpenAI a été rejointe, pour ne citer que les principaux, par les chatbots d'Anthropic (Claude), de Google DeepMind (Gemini), de Meta (LLaMA), de Mistral AI (Le Chat), d'Alibaba (Qwen), de DeepSeek (DeepSeek), de xAI (Grok)... Tous ont en commun d'être basés sur des « grands modèles de langage », ou « large language models » (LLM) en anglais (dans la suite nous parlerons surtout de « LLM »). Au total, ce sont aujourd'hui peut-être 2 milliards de personnes qui interagissent régulièrement avec un LLM grand public. Il y a 4 ans, cette technologie n'était connue que des spécialistes. Il y a 10 ans, l'idée d'une machine pouvant échanger en langage courant avec un humain et répondre, souvent correctement, à presque n'importe quelle question était de la pure science-fiction. Le bouleversement est total.

Cette technologie se distingue donc par la fulgurance de son essor, mais aussi par l'ampleur de ses effets. Cela ne bouleverse pas seulement la recherche d'information, comme Google, ou la communication et les échanges sociaux, comme Facebook, cela affecte chaque activité humaine dépendant de la cognition, c'est-à-dire... absolument toutes ! Potentiellement, il n'y a absolument rien qui échappe au regard des LLM : faire la cuisine, reformuler un CV, comprendre une leçon d'histoire, investir une partie de son épargne, réparer son toit, faire ses courses et remplir son réfrigérateur, soigner un problème de peau, synthétiser une réunion de deux heures, réorganiser son salon, écrire des lignes de code, reformuler un message délicat, créer une entreprise, se remettre en forme, traduire une documentation spécialisée, rédiger un contrat, préparer une campagne publicitaire, développer un protocole expérimental...

1.1. Une technologie impossible

Il faut vraiment prendre la mesure de cette révolution, et se rappeler de quel miracle il est question – une machine qui parle *vraiment* ! Voilà bien longtemps que nous avons des logiciels de conversation (nous reparlerons d’ELIZA un peu plus tard), et nous utilisons régulièrement des chatbots rudimentaires. Il y avait aussi déjà des systèmes plus avancés, comme l’assistant virtuel d’Amazon, Alexa, ou Siri, l’application d’Apple qui équipe les smartphones. Mais ils ne faisaient pas illusion bien longtemps. Il était impossible d’engager une conversation élaborée avec ces gadgets. Ce dont il est question depuis une poignée d’années est tout à fait différent : nous pouvons parler avec ces machines comme nous parlons à un être humain, et elles nous répondent de la même manière.

Mais l’esprit humain a cette faculté toujours étonnante de s’habituer aux changements les plus extraordinaires, et c’est aujourd’hui sans émotion que nous discutons avec ces machines, sans surprise que nous sommes témoins des prodiges dont elles sont capables.

En 1950, Alan Turing, le père fondateur de l’informatique, qui a théorisé les principes de l’ordinateur moderne, publiait un article devenu aujourd’hui un grand classique pour répondre à cette question : une machine peut-elle penser ?¹ Plutôt que d’essayer de proposer une définition de la pensée ou de l’intelligence, il a préféré contourner les questionnements métaphysiques en proposant un simple test : si une machine a plus de 30 % de chance de donner l’illusion de son humanité à un humain conversant avec elle pendant cinq minutes, alors elle mériterait d’être qualifiée de machine pensante (il ne le dit pas comme ça, mais c’est l’idée). Ce « test de Turing » a longtemps constitué le Saint Graal de l’intelligence artificielle. C’était une sorte d’horizon mythique², que les scientifiques affectaient souvent de ne pas prendre trop au sérieux³, laissant cela à un grand public (informé) avide de sensations fortes, mais qu’ils ne perdaient pas de vue. Tous ceux qui s’intéressaient un peu au sujet en étaient persuadés : si un jour une machine devait finalement réussir à le passer, ce serait l’événement du millénaire. Le monde entier ne parlerait plus que de ça, et chacun n’aurait plus à la bouche que cette nouvelle

sensationnelle : les machines intelligentes sont parmi nous ! Il n'y aurait sans doute guère que la fin du monde ou une invasion de Martiens pour rivaliser avec une actualité pareille. Après quelques annonces prématurées, le test a été passé au début de l'année 2025, l'annonce en a été faite dans un article publié le 31 mars⁴. Et personne n'en a parlé⁵.

C'est peut-être la prodigieuse difficulté de ce défi qui échappe à beaucoup de monde. C'est peut-être, à nouveau, que l'on s'habitue à *vraiment* tout. C'est peut-être la science-fiction qui nous a familiarisé depuis longtemps avec cette « possibilité ». Quoi de plus banal, pour les millions de spectateurs de Star Wars, qu'un robot capable de parler des millions de langues ?

Il faut pourtant se rappeler (ou savoir) que C-3PO n'était pas beaucoup plus scientifiquement plausible au début de ce siècle que l'hyperdrive, la téléportation ou le voyage dans le temps. Encore à la fin des années 2010, beaucoup de spécialistes de l'intelligence artificielle étaient persuadés que ces technologies étaient impossibles, les plus intrépides n'imaginaient pas qu'elles pourraient apparaître avant le milieu du XXI^e. En 2020, Luc Julia, informaticien fameux pour avoir contribué à la création de Siri, futur membre du Comité de l'intelligence artificielle générative formé par le gouvernement en 2023, souvent présenté comme un « pape de l'IA », déclare au cours d'une interview que les machines ne maîtriseront jamais le langage, car, je cite, « le langage est le propre de l'homme, et donc pas de la machine⁶ ». En 2022, Barry Smith, un mathématicien, philosophe et informaticien, et Jobst Landgrebe, un philosophe de l'intelligence artificielle, publièrent un livre au titre définitif (et rassurant), *Why Machines Will Never Rule the World: Artificial Intelligence Without Fear*, consacrant toute une section à expliquer notamment pourquoi les machines ne maîtriseraient jamais le langage humain (*Why machines will not master human language*). Le livre est paru la semaine même du lancement de ChatGPT.

Il ne faut pas se moquer : ils avaient tous de très bonnes raisons d'être sceptiques. Maîtriser le langage naturel est une opération d'une complexité qui échappe à notre entendement. C'est parfaitement impossible à codifier. Pour comprendre le langage courant, il faut être capable de comprendre une infinité de subtilités contextuelles qui semblaient

n'être accessibles qu'à l'esprit humain. Il était illusoire, et à vrai dire un peu puéril, de croire qu'un ordinateur pourrait jamais comprendre le langage courant, et *a fortiori* le parler. En 1960, un éminent mathématicien, linguiste et philosophe, Yehoshua Bar-Hillel, avait théorisé cette impossibilité dans le domaine plus restreint de la traduction automatique (mais son analyse avait évidemment une portée plus générale) en donnant l'exemple ci-dessous, aujourd'hui un grand classique des cours de linguistique informatique (nous devons évidemment donner la version originale, en plus de la traduction) :

« Little John was looking for his toy box. Finally he found it.
The box was in the pen. John was very happy⁷. »

En français, cela donne :

« Petit John cherchait sa boîte à jouets. Finalement, il l'a trouvée. La boîte était dans le parc à jouets. John était très heureux. »

C'est une traduction que, selon Bar-Hillel, seul un humain pouvait correctement faire, du fait de l'ambiguïté du mot « pen » en anglais, qui peut aussi bien signifier « stylo » que « parc à jouets ». Pour traduire correctement, il faut donc comprendre le contexte, comprendre que « pen » signifie ici « parc à jouets » et non « stylo », ce qui semblait impossible à une machine. Il faut avoir une connaissance concrète du monde où ces mots prennent sens, et l'ordinateur n'a évidemment pas cette connaissance. C'est cela qui faisait dire à Bar-Hillel que le projet d'une traduction automatisée par ordinateur était sans espoir :

« J'affirme maintenant qu'aucun programme existant ou imaginable ne permettra à un ordinateur électronique de déterminer que, dans la phrase donnée et dans le contexte donné, le mot pen a le second sens [celui de jardin d'enfants], alors que tout lecteur possédant une connaissance suffisante

de l'anglais le fera « automatiquement ». [...] Pourquoi une machine dotée d'une capacité de mémoire suffisante pour traiter un paragraphe entier à la fois, et d'un programme syntactico-sémantique qui, si nécessaire, dépasse les limites des phrases isolées jusqu'au paragraphe entier (et, pour les besoins de l'argument, jusqu'à un livre entier) [...] demeure-t-elle néanmoins incapable de déterminer le sens de pen dans notre phrase d'exemple, au sein du paragraphe donné ? L'explication est extrêmement simple [...]. Ce qui fait qu'un lecteur humain intelligent saisit ce sens avec une telle assurance [...], c'est sa connaissance du fait que les tailles relatives des « pens » au sens d'instruments d'écriture, des boîtes à jouets, et des « pens » au sens de parcs pour enfants, sont telles que, lorsque quelqu'un écrit dans des circonstances ordinaires et dans quelque chose comme le contexte donné : « The box was in the pen », il fait presque certainement référence à un parc pour enfants, et très certainement pas à un stylo. [...] Cette connaissance n'est pas à la disposition de l'ordinateur électronique, et aucun des dictionnaires ou programmes d'élimination de la polysémie ne met cette connaissance à sa disposition. >>

Ce raisonnement continuera à convaincre un demi-siècle plus tard. Le grand chercheur en sciences cognitives Douglas Hofstadter emploiera jusqu'à assez récemment cet argument, en citant explicitement Bar-Hillel, pour exprimer son scepticisme quant à la possibilité d'une véritable automatisation de la traduction (et donc de la maîtrise du langage par une machine). En 2018, c'est un véritable réquisitoire qu'il prononce en conclusion d'une conférence sur la traduction automatique (« machine translation » en anglais, ou « MT ») donnée à l'université de Cologne, en Allemagne :

« MT n'a aucune compréhension des situations, parce qu'elle n'a aucune compréhension des mots, parce qu'elle ne sait pas que les mots signifient quelque chose.

MT ne sait pas ce qu'est une colline ou une voix, ni ce qu'est le vert, ni ce qu'est la mousse.

MT ne sait rien de l'espace.

MT ne sait rien du temps.

MT ne sait rien du haut ou du bas.

MT ne sait rien du grand ou du petit.

MT ne sait pas qu'il y a des personnes.

MT ne sait pas qu'il y a des choses.

MT ne sait pas qu'il y a un monde.

MT ne sait pas que quoi que ce soit existe.

MT ne sait pas que quoi que ce soit arrive.

MT ne sait *rien*.

MT ne comprend même pas vaguement ou « plus ou moins » les textes sur lesquels elle travaille.

MT ne comprend pas.

MT est *empty* (« vide » en anglais).

Point⁸. >>

Douglas Hofstadter se veut prudent, il prend bien soin de préciser que son jugement ne vaut que pour l'état actuel des technologies. Mais le sens de son intervention est transparent : à moins d'une révolution improbable, les machines seront condamnées à maladroitement baragouiner, et à commettre des fautes qui feraient rougir un enfant de quatre ans. Le langage est le propre de l'Homme. Ce qu'il ignore en prononçant cet anathème, c'est que la révolution improbable avait eu lieu l'année d'avant, avec une invention qui allait tout changer : les « Transformers » (voir section 3).

1.2. Stupéfaction

La surprise va donc être totale. Certains, devant l'évidence des faits, feront leur *aggiornamento*. En novembre 2025, un article du *New Yorker* rapporte cet aveu de Hofstadter : « Je suis sidéré par certaines choses que font ces systèmes [...]. Cela aurait été inconcevable il n'y a même pas dix ans. [...] Ils font des choses qui ressemblent beaucoup à de la pensée [...]. On peut dire qu'ils pensent, simplement d'une manière quelque peu étrangère⁹ ». Pour d'autres, la surprise sera trop grande. Peut-être plus habités par le sentiment d'une humaine exceptionnalité, ils imagineront toutes sortes d'épicycles pour que leur univers continue à tourner rond. Mais toutes les circonvolutions conceptuelles ne détournent pas de ce fait massif : il est désormais possible de parler en langage naturel à des machines, et celles-ci répondent dans le même langage, le plus souvent de manière très cohérente. Peu importe ici qu'elles « comprennent » ou pas ce qu'elles disent, ou qu'elles aient ou non une « pensée », quoi que puissent vouloir dire ces mots (nous y reviendrons). Elles parlent¹⁰. C'est l'énormité de ce simple fait qu'il faut comprendre. L'impossible est arrivé.

La surprise est donc totale, et pas seulement pour les sceptiques ou pour tous ceux qui tenaient simplement pour acquis que le langage était le propre de l'Homme, mais aussi, même, pour ceux qui croyaient en la possibilité de machines parlantes, et jusqu'aux créateurs de ces machines. Aidan N. Gomez est l'un des coauteurs de l'article qui a changé le monde, en 2017 : *Attention is all you need*. Cet article posait les principes de fonctionnement des Transformers, cette technologie révolutionnaire qui allait conduire à la création de ChatGPT (« GPT » veut dire « Generative Pre-trained Transformer »). En 2024, il rapportait au *New Yorker* sa réaction devant les premiers résultats de la machine qu'il avait contribué à créer : « J'étais stupéfait, époustoufflé, je pensais qu'on en arriverait là dans vingt ans, vingt-cinq ans, et puis ça s'est produit¹¹ ». Andrej Karpathy, un des cofondateurs d'OpenAI et ancien directeur de la division intelligence artificielle chez Tesla, dit également sa stupéfaction face aux capacités de ChatGPT, sa propre invention : « Cela m'a époustoufflé [...]. J'ai trouvé cela miraculeux que cela ait fonctionné. C'était incroyable¹² ». Ilya

Sutskever, un des plus grands esprits de l'informatique et de l'intelligence artificielle, qui a contribué centralement à la création de l'IA générative, et qui est également cofondateur d'OpenAI, confiait en ces termes son incrédulité au micro de Dwarkesh Patel, un très fameux podcaster spécialisé sur l'IA : « Tu sais ce qui est fou ? Que tout cela soit réel¹³ ».

Les LLM ne sont pas seulement une formidable avancée technologique, c'est l'une des plus grandes découvertes scientifiques du siècle : le genre de découverte qui nous fait prendre collectivement conscience de notre ignorance. Une découverte de la même espèce et sans doute de la même importance que les observations astronomiques de Galilée au XVII^e siècle, qui révélaient la profonde méconnaissance de l'univers à cette époque, et allaient conduire à une révision complète de la cosmologie (puis, plus tard, de la physique tout entière avec Newton et Einstein). Avec les LLM, c'est notre compréhension du langage et de l'esprit humain qui va devoir être profondément réformée.

1.3. Au-delà des LLM

Et il n'y a pas que les LLM, c'est toute la technologie sous-jacente à ces machines, celle des « réseaux de neurones », qui s'est révélée d'une étrangeté et d'une puissance extraordinaire. Nous l'expliquerons plus en détail dans le chapitre 2. Dit très schématiquement, cette technologie consiste à essayer de reproduire les facultés cognitives humaines en imitant la manière dont le cerveau humain est structuré, comme un réseau de neurones donc (les neurones sont les cellules du cerveau). C'est une idée ancienne, apparue dans les années 1950, reléguée au second plan au profit de technologies qui semblaient prometteuses dans les années 1970 et 80, puis remise au goût du jour dans les années 2000. Et c'est au cours des années 2010 qu'elle a finalement tenu ses promesses, et bien au-delà.

Là encore, la surprise fut totale. Il y a un événement singulier qui la cristallise : la victoire de la machine sur l'humain au jeu de go. Plus précisément, le coup 37, le 10 mars 2016, vers 16 h 15 heure de Séoul, lors du match opposant Lee Sedol, grand champion de go, et AlphaGo, une

intelligence artificielle développée par la société DeepMind, sous la direction de Demis Hassabis. Ce n'était pas la première fois qu'un ordinateur s'imposait face à un champion humain à un jeu demandant de l'intelligence. En 1997, le légendaire champion d'échecs Garry Kasparov avait dû s'incliner devant Deep Blue, un superordinateur programmé par IBM. Mais le jeu de go semblait à l'abri de ce destin. C'est que le jeu de go est considérablement plus complexe que le jeu d'échecs, en termes de nombre de combinaisons de mouvements possibles. Il y a cent millions de milliards de quadrillards (un milliard de milliards de milliards) de quadrillards de quadrillards de quadrillards (10^{125}) de fois plus de positions possibles au jeu de go qu'au jeu d'échecs. Il est possible d'employer la « force brute » au jeu d'échecs pour battre un champion, en calculant à chaque seconde des millions de stratégies possibles. C'est ce que faisait Deep Blue. Cette manière de faire, finalement très peu intelligente, est inenvisageable avec le jeu de go. Avec le go, il faut de l'intuition, ce qui mettait, presque personne n'en doutait, la machine hors-jeu, peut-être pour toujours, sans doute pour au moins un ou deux siècles¹⁴.

Entre le 8 et le 15 mars 2016, un grand match de démonstration est organisé entre Lee Sedol et AlphaGo. Lee Sedol perd la première partie, le 9 mars. C'était déjà inattendu. En fait, c'était un véritable tremblement de terre, qui a fait la une des journaux coréens, et a été commenté dans le monde entier. Mais c'est vraiment lors de la seconde partie que des bras vont tomber et des mâchoires se décrocher. Il est 16 h 12, le 10 mars. La partie a commencé depuis plus de trois heures (elles commencent toutes à 13 h), et Lee Sedol décide de prendre une pause, pour aller fumer une cigarette sur la terrasse de l'hôtel où se joue l'affrontement. C'est au tour d'AlphaGo de jouer, et l'assistant chargé de placer les pions pour la machine en dépose un sur le tablier (le « goban »). C'est le mouvement 37. Un murmure parcourt la salle où se tient le public, les commentateurs s'exclament et s'étonnent : « un mouvement impensable ». Dans la salle de contrôle, l'équipe de DeepMind s'affaire à comprendre ce qui se passe. Lee Sedol revient, s'assoit à la table, regarde le goban, et son monde change : « Je pensais qu'AlphaGo reposait sur des calculs probabilistes et qu'il n'était qu'une machine. Mais quand j'ai vu ce coup, j'ai changé

d'avis. Assurément, AlphaGo est créatif. Ce coup était vraiment créatif et beau¹⁵ ».

Ce jour-là, cet instant-là, l'humanité apprenait qu'il y avait dans ces réseaux de neurones plus qu'une version modernisée des automates de Vaucanson, plus que des rouages électroniques. Il y avait peut-être un peu du mystère de l'humain. Et cette part mystérieuse n'allait faire que croître, et elle continue à croître aujourd'hui.

Ces réseaux de neurones ont donné les LLM, ces machines qui travaillent le texte, mais il y a aussi toutes les autres « IA génératives » qui ne produisent pas du texte, mais des images, des vidéos, de la parole, du son ou de la musique. Ces machines vont radicalement transformer – transforment déjà radicalement – les domaines de la création artistique, là où la créativité humaine était encore seule en scène au début des années 2020.

Et au-delà des IA génératives, les réseaux de neurones ont donné une infinie variété d'IA spécialisées dans divers domaines. La même équipe qui a créé AlphaGo créera ensuite AlphaFold, une IA spécialisée dans l'analyse de la structure des protéines, ces molécules essentielles au bon fonctionnement des organismes vivants. Il est très simple de déterminer la composition d'une protéine, mais incroyablement difficile d'en comprendre la structure. Il faut des mois, des années de travail à des équipes entières de chercheurs pour déterminer la structure d'une seule protéine. Or la connaissance de cette structure est essentielle pour en comprendre les effets médicaux. C'est donc un enjeu sanitaire et économique de première importance. En 70 ans, depuis que la science dispose des moyens nécessaires à cette tâche, jusqu'à la fin des années 2010, seules environ 100 000 structures de protéines ont été trouvées, sur les milliards qui existent. Le 28 juillet 2022, l'équipe de DeepMind rendait publiques les structures d'environ 200 millions de protéines. AlphaFold les avait déterminées en quelques mois. En 2024, Demis Hassabis partageait le prix Nobel de chimie avec John Jumper, le chimiste qui l'avait aidé à créer AlphaFold. Au moment où sont écrites ces lignes, personne n'est encore capable de prendre la mesure des progrès médicaux que permettra cette révolution, une parmi d'autres.

C'est un nouveau monde qui se lève, et par bien des aspects il sera très différent du précédent. Même en faisant l'hypothèse peu plausible que les progrès de ces technologies cesseront demain, et que nous en resterons aux IA telles qu'elles sont aujourd'hui, tout va changer. À commencer par nos compétences, et c'est de cela qu'il sera question ici.

1.4. Le deuil de l'ancien monde

À chaque révolution technologique, ce sont des compétences qui disparaissent, et d'autres qui apparaissent. Et souvent de manière très inattendue.

La même histoire se répétera avec les IA génératives : nous allons perdre et nous allons gagner. Et il serait sans doute vain d'essayer de faire le bilan de ces gains et de ces pertes. Dans ce nouveau monde, ils n'auront plus le même sens qu'aujourd'hui. L'idée de ne plus savoir allumer un feu avec deux bouts de bois, l'un dur et l'autre tendre, aurait sans doute terrorisé nos aïeux du 400^e siècle av. J.-C. Mais que nous importe aujourd'hui ?

Mais, et c'est bien normal, ce tournant nous effraie tout de même. La souffrance d'une perte concrète n'est pas effacée par l'espoir d'un gain encore abstrait. Et puis, les économistes l'ont cent fois démontré, nous valorisons moins les gains que l'évitement des pertes.

Cette peur n'est pas nouvelle. Platon, dans le *Phèdre* (274c-275b), rapporte par la bouche de Socrate le mythe des origines de l'une des premières technologies cognitives : l'écriture. Son inventeur est un dieu, Theuth. Et Theuth est tout fier d'aller présenter sa trouvaille au pharaon, Thamous. Mais Thamous ne voit pas ça d'un bon œil. L'écriture, répond Thamous à Theuth,

« développera l'oubli dans les âmes de ceux qui l'auront acquise, par la négligence de la mémoire ; se fiant à l'écrit, c'est du dehors, par des caractères étrangers, et non du dedans, et grâce à l'effort personnel, qu'on rappellera ses

souvenirs. Tu n'as donc pas trouvé un remède pour fortifier la mémoire, mais pour aider à se souvenir. Quant à la science, tu en fournis seulement le semblant à tes élèves, et non pas la réalité. Car, après avoir beaucoup appris dans les livres sans recevoir d'enseignement, ils auront l'air d'être très savants, et seront la plupart du temps dépourvus de jugement, insupportables de surcroît parce qu'ils auront l'apparence d'être savants, sans l'être. >>

Le mythe de Theuth est souvent convoqué pour se rassurer. Les craintes de Thamous étaient manifestement exagérées. Notre univers de l'écrit n'a pas fait de nous ces oublieux sans substance que dépeignait le pharaon. Platon, Aristote, Archimède, Ptolémée, Avicenne, Copernic, Galilée, Newton, tous ces géants juchés sur les épaules d'autres géants, n'avaient pas seulement l'apparence d'être savants, ils l'étaient, sans doute plus que les meilleurs esprits de la cour de Thamous, qui n'avaient pas le secours de l'écriture pour avancer dans la connaissance de l'univers. Alors pourquoi faudrait-il s'inquiéter, quand de pareils discours nous sont resservis à propos des dernières technologies cognitives ? Dormons tranquilles, et laissons les Philippulus de toutes espèces à leurs prophéties apocalyptiques.

Mais il faut peut-être comprendre autrement la leçon socratique : Thamous avait raison ! Au moins partiellement. Oui, l'écriture nous a fait perdre quelque chose. Qui aujourd'hui saurait réciter l'Iliade d'une traite ? Notre mémoire n'est sans doute pas la même que celle de nos aïeux. Et nous l'avons accepté.

Nous devons accepter que certaines compétences disparaissent avec les révolutions technologiques. Et il ne s'agit pas de dire que ces compétences disparues seront plus ou moins compensées par d'autres compétences qui vont apparaître. Nous l'avons dit, cela n'aurait sans doute pas beaucoup de sens. Il s'agit de faire son deuil de ces compétences en faisant le deuil du monde qui leur donnait sens. Car ces compétences qui s'évanouissent disparaissent avec les mondes qui les rendaient particulièrement désirables. Cela n'aurait aucun sens d'apprendre aujourd'hui l'Iliade par cœur. Ce serait seulement un exploit cognitif un peu bizarre, à ranger au rayon du

genre de records baroques que collectionne le livre Guinness. Cela n'aurait pas plus de sens, par exemple, de savoir se servir d'une pascaline, d'un boulier ou d'une règle à calculer, car toutes ces compétences appartiennent à des mondes révolus.

Il n'est évidemment pas question de faire le deuil de toutes les compétences affectées par ces changements. Les révolutions technologiques peuvent en rendre certaines moins centrales sans les faire disparaître complètement. L'écriture n'a évidemment pas fait disparaître la nécessité de mémoriser, elle a seulement rendu cette compétence généralement moins pressante. La calculatrice n'a pas fait disparaître la nécessité de savoir calculer de tête, elle a seulement autorisé, dans la plupart des circonstances, une moindre dextérité dans l'exercice de cette faculté. Mais il faut toujours en être capable.

Et parfois, ces compétences mises au second plan vont devenir des facteurs de distinction, précisément parce qu'elles ne sont plus des obligations universelles. Dans le domaine de la finance, être doué en calcul mental est un atout considérable, et fait la différence entre ceux qui peuvent affronter une salle de marché et les autres. Ce serait une très mauvaise idée, pour un étudiant aspirant aux métiers de la finance, de s'en remettre exclusivement à sa calculatrice pour passer ses examens de mathématiques. De la même manière, un dirigeant qui négligerait d'entraîner sa mémoire, en se reposant toujours sur l'écrit, sera pris fort au dépourvu dès que la pression des événements commandera de prendre très vite des séquences entières de décisions sans lui laisser le temps de consulter ses dossiers. De la même manière, les LLM ne feront sans doute pas disparaître la nécessité de savoir écrire. Au contraire, elles feront peut-être de cette compétence un talent aristocratique, qui distinguera les meilleurs.

Ceux qui croient pouvoir se reposer sur la technologie et négligent de cultiver ce genre de talents paieront le prix de leur insouciance. Ce sera très bientôt une métacompétence indispensable que de savoir reconnaître les compétences que l'IA générative va faire disparaître, celles qu'elle transformera, et celles qui deviendront l'apanage d'une élite. C'est l'ambition de ce livre d'aider à développer cette métacompétence.

Mais ce n'est pas tout. Car, nous le disions au début de cette section, si certaines compétences vont disparaître et d'autres vont changer, d'autres encore vont apparaître. Il va falloir essayer de les anticiper, et c'est un exercice périlleux qui s'annonce, car les chemins que les révolutions technologiques font prendre au monde sont souvent sinueux, et leurs destinations toujours surprenantes.

1.5. Se préparer au nouveau monde

Il y a 2 700 ans environ, quelqu'un en Grèce a l'idée d'ajouter une seconde poignée à son bouclier de combat. Il invente ainsi l'hoplite, ce bouclier doté d'une double poignée (le *porpax*, pour passer l'avant-bras, et l'*antilabe*, pour la main). Cette innovation permet de combattre en formation serrée, la phalange, et transforme radicalement l'art de la guerre. Ce n'est plus une affaire de héros – Achille, Ulysse, ou Diomède sont renvoyés à leur passé glorieux – c'est une affaire collective. Chacun désormais va pouvoir défendre sa cité, tous à égalité au sein de la phalange, chaque soldat, riche ou pauvre, protégeant son voisin. C'est la révolution hoplitique¹⁶, qui va donner naissance à la démocratie. Car le pauvre qui risque sa vie à égalité avec son voisin plus riche va réclamer les mêmes droits, la même place dans la cité. C'est ainsi que d'une innovation technologique sont nées, outre de nouvelles compétences martiales, les compétences citoyennes¹⁷.

Avec ce livre, nous ne prétendons pas voir aussi loin, mais nous ne perdrons pas de vue que les conséquences des IA génératives pourraient bien être aussi surprenantes que celles du *porpax*, et qu'il faut garder l'esprit ouvert et ne pas se laisser prendre au sortilège des évidences immédiates. La littérature florissante sur l'IA, générative ou pas, est trop souvent myope, obnubilée par les apparentes conséquences immédiates et par l'extrapolation des tendances du moment. Les étudiants utilisent massivement ChatGPT pour écrire leurs devoirs ? C'est l'écriture qui va disparaître. Des juristes font produire aux machines des avis douteux ? C'est le droit qui va s'effondrer. Chacun de nous demande son avis à

une IA avant de prendre la moindre décision ? C'est notre capacité collective à raisonner, et avec elle à construire une vie collective rationnelle, qui va s'évanouir. Ce sont des fables faciles, qui attirent un public friand d'émotions fortes et de peurs à bon marché, qui font vendre et qui font parler, mais qui n'aident pas à comprendre les véritables enjeux. Il faut bien sûr s'inquiéter des conséquences nocives de l'IA, car il y en a et il y en aura évidemment, mais cela doit se faire de manière intelligente, en portant un regard objectif sur les IA génératives, en considérant ce qui n'apparaît pas au premier regard, et en réfléchissant aux conséquences futures, même inattendues. C'est de cette manière que nous pourrions anticiper ces nouvelles compétences qui viennent et qui viendront avec ces machines. C'est l'objectif de ce livre.

Conçu comme le compagnon d'un cours suivi par l'ensemble des étudiants de ESCP Business School, mais écrit pour tout praticien ou usager de ces technologies, cet ouvrage entend offrir des repères durables plutôt qu'un catalogue de recettes ou une énième critique alarmiste. Car si les innovations dans le domaine des IA arrivent maintenant à un rythme difficile à suivre, les fondamentaux technologiques de ces IA (LLM, RL, attention, embeddings, etc.), qui déterminent leurs forces et leurs faiblesses, ont été inventés dans les années 2010 et seront sans doute encore au cœur de ces machines dans les années 2030.

Notre ambition est double. D'une part, fournir une compréhension rigoureuse – mais accessible – de ces bases technologiques des IA génératives textuelles, et analyser ce qu'elles impliquent en termes de capacités cognitives, au-delà des lieux communs qui encombreront les débats sur ces IA. D'autre part, s'appuyer sur ce diagnostic pour penser notre entrée dans ce nouveau monde qui s'annonce. Dans cette nouvelle « jungle » numérique, nous voulons proposer un véritable guide de survie intellectuelle, dont la valeur résidera moins dans la maîtrise d'outils éphémères que dans l'exercice éclairé de certaines vertus et compétences.

Nous refusons donc le catastrophisme comme l'enthousiasme naïf. Ces systèmes demeurent certes, techniquement parlant, des « machines à bullshit » : ils sont structurellement incapables de distinguer par eux-mêmes le vrai du faux. Mais ils rendent également possible, à coût marginal

quasi nul, la génération d'idées, de prototypes ou d'analyses qu'il aurait été difficile d'envisager il y a cinq ans. L'enjeu n'est donc pas de décider s'il faut les utiliser, mais de déterminer comment, quand, sous quelles conditions et évidemment avec quelles dispositions intellectuelles.

Ce livre sera constitué, outre l'introduction, de six chapitres. Le second chapitre expliquera en des termes précis mais accessibles ce qu'est l'IA en général, et l'IA « connexionniste » en particulier (basée sur des « réseaux de neurones artificiels »), et il introduira la notion d'IA générative. Nous rentrerons un peu dans la technique, car il est impossible de comprendre les enjeux pratiques dont il sera question dans la suite sans comprendre les grandes lignes du fonctionnement de ces technologies. Le troisième chapitre explorera plus en détail les technologies au cœur de ce qui est sans doute l'IA générative la plus révolutionnaire de toutes : les chatbot basés sur des grands modèles de langage. Là encore, la compréhension détaillée du fonctionnement de ces innovations est absolument nécessaire pour savoir quelles compétences acquérir et quelles vertus cultiver. Avec le quatrième chapitre, nous reviendrons sur toute une série de lieux communs qui le plus souvent obscurcissent la compréhension de ces machines (les IA génératives ne sont pas « intelligentes », elles ne « comprennent » rien, etc.). C'est là à un travail de déblayage que nous nous livrerons, avant d'en venir à leurs véritables limites structurelles, qui seront exposées dans le cinquième chapitre. Nous y expliquerons notamment pourquoi ces instruments sont, dans un sens tout à fait technique et objectif, sans le moindre jugement de valeur, des « machines à bullshit », ou de « gentils psychopathes ». Dans le sixième chapitre, le plus long, nous donnerons la parole aux professionnels de divers domaines, pour essayer, à partir de leurs pratiques, de comprendre ce que seront ces nouvelles compétences que vont faire apparaître les IA génératives. Enfin, en conclusion, nous proposerons aux lecteurs une « grille de diagnostic » permettant à chacun d'identifier les compétences qui doivent être développées selon les usages envisagés des IA génératives.

2

L'IA GÉNÉRATIVE, QU'EST-CE QUE C'EST ?



Une IA générative, c'est une IA qui génère du texte (notamment de la prose, mais aussi du code, des données structurées comme des tableaux ou des productions symboliques plus abstraites), de l'image (animée ou non) ou du son (notamment des paroles ou de la musique). Ce n'est pas le cas de toutes les IA. Beaucoup traitent des données, résolvent des problèmes, reconnaissent des formes, repèrent différentes sortes de régularités sans générer ni texte, ni image, ni son.

Mais au fait, avant d'aller voir plus en détail ce qu'est une IA générative, qu'est-ce qu'une IA ?

2.1. Définir l'IA

Pour commencer, l'IA est un domaine de recherche. C'est par métonymie que l'on parle d'IA à propos des choses qui sont produites par les scientifiques travaillant dans ce domaine de recherche, ou par les industriels appliquant les résultats de leurs travaux de recherche. Mais cela ne nous renseigne pas beaucoup. Ces choses qui sont produites dans ou à partir du domaine de l'intelligence artificielle, comment les définir ?

Malheureusement, il n'y a pas vraiment de définition satisfaisante.

Prenons par exemple celle proposée par l'Union européenne pour son « AI Act », le règlement (UE) 2024/1689, dans son article 3 (1) :

« un « système d'IA » est un système automatisé conçu pour fonctionner avec différents niveaux d'autonomie, pouvant s'adapter après déploiement, et qui, pour des objectifs explicites ou implicites, déduit à partir des entrées comment générer des sorties (notamment prédictions, contenu, recommandations, décisions) pouvant influencer des environnements physiques ou virtuels. »

À ce compte-là, un thermostat est une IA, une soupape de cocotte-minute également : c'est automatique, autonome, ça peut s'adapter, ça

a un objectif explicite (réguler la température ou la pression), ça génère une décision (modifier la température, libérer de la vapeur) à partir des données d'entrée (la température d'une pièce, la pression de la cocotte). Ça marche aussi avec le régulateur à boules de Watt, le métier à tisser Jacquard (ce qui dans ce cas n'est pas complètement idiot, puisqu'il est souvent mentionné dans l'histoire de l'informatique), ou... la chasse d'eau à flotteur. Manifestement, cette définition embrasse trop large.

La CNIL propose une autre définition, sibylline : « un procédé logique et automatisé reposant généralement sur un algorithme et en mesure de réaliser des tâches bien définies¹ ». Là encore, un thermostat ferait l'affaire, ou un réveille-matin, ou une montre. La définition proposée par le Parlement européen semble plus spécifique : « L'IA désigne la possibilité pour une machine de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité² ». Clairement, cette fois-ci, un thermostat ne répond pas à ces critères. Le problème, c'est qu'on ne sait pas bien ce que sont ces « comportements liés aux humains » qui font d'une machine une IA. Extraire une racine carrée, est-ce un « comportement lié aux humains » ? Si oui, alors une calculatrice est déjà une IA. Le même problème se pose avec la définition proposée sur Wikipédia : « L'intelligence artificielle (IA) est l'ensemble des systèmes informatiques capables d'effectuer des tâches typiquement associées à l'intelligence, telles que l'apprentissage, le raisonnement, la résolution de problèmes, la perception ou la prise de décision³ ». Là encore, beaucoup de systèmes qui ne sont généralement pas tenus pour des intelligences artificielles, comme le pilote automatique d'un avion (qui prend des décisions), pourraient satisfaire cette définition.

Pour mieux comprendre ce qu'est une IA, il vaut mieux finalement revenir à la définition du domaine de recherche éponyme. Sa date et son lieu de naissance officiels sont très précis : le 18 juin 1956 à Hanover (New Hampshire, États-Unis), au Dartmouth College. Bien sûr, on peut faire remonter la notion ou le projet d'intelligence artificielle à plus loin, et même beaucoup plus loin, avec par exemple le mythe des automates d'Héphaïstos, celui de Pygmalion et Galatée, ou évidemment celui du Golem. Sans revenir à ces très vieilles histoires, l'idée était bien dans l'air depuis

quelques années, depuis la création des premiers ordinateurs électroniques, dans les années 1940. Turing avait déjà publié son article de 1950 sur les « machines qui pensent ». Les bases théoriques des futurs IA avaient été posées dès 1943 dans un article introduisant la notion de neurones artificiels (voir section 2.3). Mais c'est ce jour-là, le 18 juin 1956 (ou à peu près, il y a des incertitudes sur cette date), à cet endroit-là, au Dartmouth College de Hanover, que l'IA comme domaine de recherche a été portée sur les fonts baptismaux. C'est là que s'est tenue, pendant quelques semaines, une très fameuse conférence (la « conférence de Dartmouth ») où il sera explicitement question, pour la première fois, d'intelligence artificielle. L'argument de la conférence, rédigé l'année d'avant, le 31 août 1955, annonce clairement le programme :

« Nous proposons qu'une étude sur l'intelligence artificielle, d'une durée de deux mois et mobilisant dix personnes, soit menée durant l'été 1956 au Dartmouth College, à Hanover (New Hampshire). L'étude doit se dérouler sur la base de la conjecture selon laquelle tout aspect de l'apprentissage, ou toute autre caractéristique de l'intelligence, peut en principe être décrit avec une précision telle qu'il soit possible de construire une machine capable de le simuler. On cherchera à déterminer comment faire en sorte que des machines utilisent le langage, forment des abstractions et des concepts, résolvent des types de problèmes aujourd'hui réservés aux humains, et s'améliorent elles-mêmes. »

C'est l'occasion de tordre le cou à une rumeur insistante, qui prétend que le mot « intelligence », dans « intelligence artificielle », aurait à l'origine signifié « renseignement », qui est l'un des sens du mot anglais « intelligence » (comme dans CIA, « Central Intelligence Agency »). Non, c'est tout à fait clair, l'intention des fondateurs du domaine n'avait rien à voir avec le renseignement, et tout avec les plus hautes facultés cognitives de l'humain. Ce qu'ils avaient en tête, quoique cela restait encore vague, était assez clair : donner une intelligence humaine à des machines.

Non pas seulement extraire une racine carrée, mais parler, inventer, comprendre, apprendre. Allen Newell, qui a participé à la conférence de Dartmouth et est donc l'un des fondateurs du domaine, proposait cette définition en 1973 :

« L'IA est un domaine dédié à construire des artefacts intelligents, où "intelligent" est opérationnalisé via des tests d'intelligence (comme le Wechsler Adult Intelligence Scale) et d'autres tests d'aptitudes mentales (incluant par exemple des tests de raisonnement mécanique, de créativité, etc.)⁴. »

Cette définition du domaine de l'IA permet de comprendre de quoi il est vraiment question lorsque l'on parle d'IA : de machines qui *visent* à devenir aussi intelligentes que les humains. Il ne s'agit donc pas de dire qu'une IA serait par définition une machine capable de passer des tests d'intelligence. Une telle définition, au contraire de celles déjà mentionnées, serait beaucoup trop restrictive. L'immense majorité des IA n'ont pas ce genre de faculté. En fait, toutes les IA d'avant la révolution des LLM en étaient parfaitement incapables. Mais toutes ont joué un rôle dans la réalisation de ce but, et c'est cela qui les définit. Une IA, c'est une machine qui conforte le projet d'intelligence artificielle. Chaque IA créée depuis 70 ans a démontré ou confirmé qu'une faculté cognitive avancée pouvait être maîtrisée par une machine, depuis la capacité à percevoir et reconnaître des formes, avec le « perceptron » des années 1950, jusqu'au langage, avec les LLM actuels.

Ce programme est un pari philosophique, car beaucoup doutent qu'une machine puisse être véritablement dotée de facultés « cognitives », qu'elle puisse « penser ». Une machine peut détecter des signaux lumineux et produire en retour d'autres signaux, mais ne pourrait pas « voir » ni encore moins « comprendre » ce qu'elle « voit ». Une machine peut émettre des sons qu'un humain peut reconnaître comme du langage, mais ne pourrait pas « parler ». Dans le cas des machines, ces mots ne seraient employés que par abus de langage. Il faudrait un « quelque chose en plus »,

un « souffle vital », une « âme » pour animer ces misérables amas de matière et les rendre vraiment capables de ce que l'humain peut faire.

C'est la position des « dualistes », qui considèrent que l'esprit se distingue du corps, qu'il s'agit de deux choses différentes. C'est une position qui ne convainc pas grand monde parmi les scientifiques, et pas beaucoup plus parmi les philosophes, car ce genre d'entités mystérieuses (l'âme, l'esprit, le souffle vital...) est insaisissable presque par définition, et l'économie générale des doctrines dualistes ressemble beaucoup trop à une pétition de principe confortable. Sans compter que ça sent fort l'encens... Quoi qu'il en soit, l'IA ne repose pas sur une conception dualiste de l'esprit, mais sur une conception « naturaliste » : l'esprit est un phénomène naturel, rien de plus qu'un certain arrangement dynamique de la matière. Rien ne s'oppose donc à ce que ce phénomène apparaisse dans une autre sorte d'arrangement de matière. Il faut reconnaître que cette position n'est pas beaucoup plus confortable que celle des dualistes, car la manière dont l'intelligence et les autres facultés de l'esprit peuvent sortir de ce tas de matière n'est pas beaucoup moins mystérieuse que l'Immaculée Conception. Les scientifiques et les philosophes matérialistes disent que l'esprit et les facultés cognitives sont des propriétés « émergentes » de la matière, ce qui est une façon savante d'admettre son ignorance. Mais au moins, cela permet de partir de quelque chose de concret : la machine.

Mais il ne suffit pas d'être naturaliste, il faut aussi être « fonctionnaliste » : l'intelligence est un ensemble de fonctions (résoudre des problèmes complexes, s'adapter à des environnements changeants, manipuler mentalement des concepts abstraits...), et c'est cela qui importe. Beaucoup de celles et ceux qui critiquent l'IA ont une approche anti-fonctionnaliste : ils ou elles pensent que le substrat de l'intelligence importe, et que les IA ne feront donc jamais qu'imiter cette faculté. La véritable intelligence ne peut surgir que d'un organisme vivant. Les anti-fonctionnalistes ne pensent pas qu'il faut une « âme » ou quelque chose comme ça. On peut refuser le fonctionnalisme en restant naturaliste (on peut aussi être fonctionnaliste sans être naturaliste, croire que la fonction de l'esprit est réalisée par autre chose que des phénomènes naturels, mais c'est beaucoup moins courant). Ils ou elles pensent que la matière dont sont constitués les êtres intelligents

détermine cette intelligence. Il faut du biologique, pas du minéral, du carbone, pas du silicium, du vivant, pas de l'inerte. Il faut certaines réactions chimiques, certaines molécules, certains états neurophysiologiques.

Les fonctionnalistes pensent que tout ça, ce sont des détails, qu'une intelligence est pleinement réalisée du moment que ses fonctions sont réalisées, peu importe par quoi. Dirait-on d'une machine qui donne l'heure et qui peut se porter au poignet qu'elle « imite » les montres ? Non, il s'agirait d'une montre. Une machine qui réalise les fonctions d'une montre (donner l'heure et pouvoir être portée au poignet) *est* une montre. Le fonctionnaliste croit que l'intelligence est d'une même espèce que la montre : l'imiter, c'est la réaliser. Réaliser ses fonctions, c'est la réaliser tout entière. Bien sûr, dans le cas de l'intelligence, personne ne sait exactement ce que sont les fonctions en question, et c'est l'objet de controverses incessantes. Mais l'idée est là.

2.2. Les paradigmes de l'IA

Pour arriver à leur fin, les fondateurs du domaine de l'IA envisagèrent deux voies possibles. La première consistait à imiter le raisonnement humain. La seconde à imiter le fonctionnement du cerveau humain. À ces deux voies correspondent les deux principaux « paradigmes » de l'IA : le paradigme symbolique et le paradigme connexionniste (il y a d'autres approches possibles, bayésiennes notamment, mais ne compliquons pas une histoire déjà bien compliquée). Les IA génératives modernes sont essentiellement issues du paradigme connexionniste, mais il faut dire quelques mots du paradigme symbolique, qui a toujours un rôle à jouer.

« Imiter le raisonnement humain », cela veut dire implémenter dans la machine les règles de ce raisonnement. Il y a fort longtemps que de grands esprits ont essayé de comprendre ces règles. Cela a vraiment commencé avec Aristote, qui a formalisé la logique en formulant ses grands principes (comme la non-contradiction) et en théorisant les syllogismes (« Socrate est un homme, tous les hommes sont mortels, donc Socrate est mortel »). Les savants du Moyen Âge ont suivi le chemin tracé par Aristote, notamment

les philosophes scolastiques, et au XIX^e puis au XX^e siècle, des logiciens, des philosophes et des mathématiciens (Gottlob Frege, David Hilbert, Bertrand Russell, Kurt Gödel, Alonzo Church, Alan Turing) parachevèrent cet ouvrage en définissant les termes d'une complète mécanisation du raisonnement, en montrant sous quelles conditions il pouvait être vu comme une série d'étapes mécaniques qu'une machine pouvait exécuter. Cela a par exemple donné les logiciels de calcul formels, qui permettent de résoudre toutes sortes de calculs, même avancés, la démonstration automatique de théorèmes, ou les systèmes de planification automatique.

Cette approche de l'intelligence artificielle consistant à formaliser le raisonnement humain pour l'implémenter dans une machine, sous la forme d'une série d'instructions, a notamment donné les « systèmes experts », qui ont dominé le domaine de l'IA entre 1970 et 1990 environ. Le principe était le suivant : on demandait à des experts d'un domaine particulier, par exemple la médecine, d'exposer tout leur savoir à un informaticien qui le codait et l'implémentait dans un ordinateur. Par exemple, si un patient a de la fièvre, des maux de tête et une raideur de la nuque, alors il y a 80 % de chances qu'il s'agisse d'une méningite. Les systèmes experts sont ainsi de gigantesques algorithmes, d'immenses ensembles d'instructions logiques du type « si x alors y ».

Bien que ces systèmes aient obtenu des résultats impressionnants dans certains domaines très particuliers (comme le diagnostic d'infections sanguines avec le système MYCIN), ils se sont heurtés à un obstacle majeur : il est impossible de codifier sous forme de règles explicites l'immensité des connaissances tacites et intuitives que nous utilisons au quotidien pour agir dans le monde. Il est notamment impossible de codifier parfaitement le langage humain. Les règles de grammaire et les connaissances lexicales sont très loin de suffire à expliquer ses innombrables subtilités, comme le rappelait Bar-Hillel en 1960. Le paradigme symbolique a donc laissé la place au paradigme connexionniste, qui repose sur une intuition tout à fait différente : pour qu'une machine soit intelligente, elle doit imiter le fonctionnement du cerveau humain.

Le cerveau humain est essentiellement composé d'environ 86 milliards de cellules appelées « neurones » (« essentiellement », car il faudrait aussi compter les cellules gliales, qui sont en nombre comparable mais

dont le rôle dans le traitement de l'information est moins central), et chacun de ces neurones est connecté jusqu'à environ 10000 autres via des « synapses », des jonctions intercellulaires par lesquelles transitent des signaux électrochimiques. Lorsque ces signaux sont reçus par un neurone via ces synapses, il les traite et peut renvoyer en retour un autre signal à d'autres neurones via son « axone ».

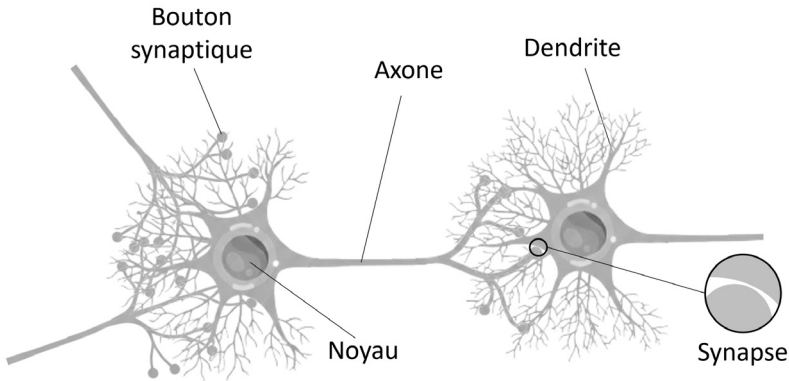


Figure 1. Schéma de neurones interconnectées⁵.

Le cerveau est donc un gigantesque réseau de neurones, et c'est cette cascade sans fin de réactions électrochimiques qui permet de peindre la Joconde, de concevoir la relativité générale, d'écrire *À la recherche du temps perdu* ou de faire la vaisselle et le ménage. Le pari du paradigme connexionniste, c'est qu'une machine fonctionnant de manière analogue, avec le même type de structure, pourrait réaliser ce genre de prodiges (on ne réalise généralement pas à quel point faire le ménage ou la vaisselle est un prodige dont les machines sont encore largement incapables, alors qu'elles peuvent sans peine résoudre presque n'importe quelle équation). Et c'est encore un pari philosophique, car tout le monde ne croit pas que le génie humain tient tout entier dans de gros tas de neurones.

C'est un pari plus spécifique que le naturalisme fonctionnaliste du programme général d'intelligence artificielle. C'est le pari de l'émergentisme. L'idée est la suivante : construire des réseaux de neurones artificiels suffisamment grands et complexes pour que puissent en émerger des facultés cognitives avancées.

Il faut distinguer deux sortes d'émergentisme : d'une part l'émergentisme « métaphysique », ou « fort », qui défend que les propriétés émergentes macroscopiques, en l'occurrence les facultés cognitives, sont « ontologiquement » irréductibles aux propriétés microscopiques, qu'elles ont une existence indépendante du fait de leur nature profonde ; d'autre part l'émergentisme « épistémologique », ou « méthodologique », ou « faible », qui soutient que les propriétés émergentes ne sont pas d'une nature différente des propriétés microscopiques, qu'il n'y a pas de rupture ontologique entre les niveaux micro et les niveaux macro, mais qu'il est impossible ou très difficile de prédire ou d'expliquer les unes à partir des autres : la rupture est épistémologique (en lien avec la connaissance), et non plus métaphysique (en lien avec la réalité).

L'émergentisme des connexionnistes est de ce second type, mais c'est un émergentisme faible particulier : les connexionnistes ne disent pas nécessairement qu'il est impossible de comprendre les facultés de haut niveau à partir des propriétés de bas niveau (en général, ils reconnaissent que c'est très difficile, et nous y reviendrons), ils disent que ce n'est pas nécessaire, que l'on peut très bien se passer de ces connaissances pour faire émerger ces facultés d'une machine. On ne sait pas pourquoi ça émerge, mais c'est pas grave. Il s'agit d'une approche du bas (le réseau de neurones) vers le haut (les facultés cognitives qui émergent du réseau), contrairement à l'approche symbolique, qui va du haut (les règles et normes abstraites de la raison) vers le bas (les facultés cognitives qui sont le résultat de l'application de ces règles).

Le pari n'est pas seulement philosophique, il est technique et économique : à force de faire grossir des réseaux de neurones artificiels, avec des ordinateurs toujours plus puissants, des sources de données toujours plus imposantes, on finira par voir apparaître des comportements cognitifs avancés qui pourront être employés à grande échelle par les organisations publiques ou privées ou par les particuliers. Cette application du « bigger is better » aux IA est au départ des « scaling laws » qui organisent aujourd'hui l'industrie des IA génératives (voir la section 3.6).

Cette approche est aujourd'hui victorieuse, elle domine largement l'IA, mais ça n'a pas toujours été simple. La conférence de Dartmouth

était largement dominée par les tenants de l'approche symbolique, et si le programme connexionniste avait sa place dans la communauté naissante de l'IA, cette approche restait minoritaire et est devenue complètement marginale après 1969 et la publication par les grands informaticiens Marvin Minsky et Seymour Papert de critiques dévastatrices des réseaux de neurones⁶, et plus spécifiquement du perceptron inventé 12 ans plus tôt par Frank Rosenblatt. C'est le premier « hiver de l'IA », dont l'IA sortira avec le développement des systèmes experts, et il faudra attendre l'échec de ces IA symboliques, incapables de tenir leurs promesses, et donc le second « hiver de l'IA », pour que les réseaux de neurones puissent venir sur le devant de la scène, dans les années 1990. Aujourd'hui, ils n'occupent plus seulement le devant de la scène, mais aussi l'arrière, les côtés et le centre. Dans le monde académique, c'est plus de 95 % des publications qui concernent le paradigme connexionniste.

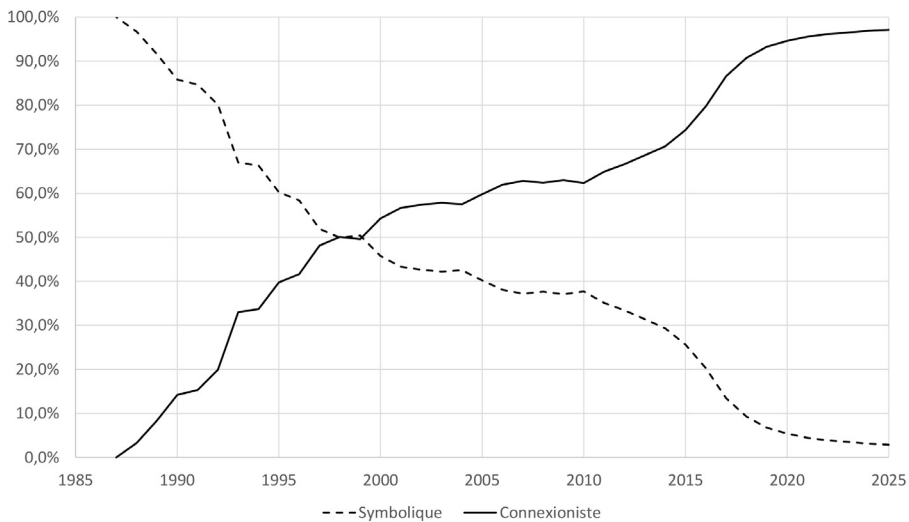


Figure 2. Parts respectives des publications issues des paradigmes connexionniste et symbolique dans la production académique du domaine de l'IA⁷.

En quarante ans, le rapport de domination s'est totalement inversé. Aujourd'hui, l'IA, ce sont les réseaux de neurones. Les « algorithmes » de recommandation de Netflix, d'Amazon ou de YouTube : des réseaux

de neurones. La reconnaissance faciale de votre iPhone ou la vidéosurveillance dans les rues : encore des réseaux de neurones. Votre boîte e-mail qui envoie les messages indésirables dans le spam : toujours des réseaux de neurones. La quasi-totalité des IA que nous utilisons au quotidien, parfois sans le savoir, sont des réseaux de neurones. Pour comprendre l'IA d'aujourd'hui, c'est cette technologie qu'il faut comprendre.

2.3. Réseaux de neurones

Pour construire un réseau de neurones artificiels, il faut d'abord modéliser les réseaux de neurones naturels, en produire une représentation abstraite, mathématique, débarrassée des innombrables complexités biologiques qui permettent à un cerveau de fonctionner concrètement mais qui n'expliquent pas centralement sa fonction principale : produire de l'intelligence. Cette démarche de modélisation est analogue à celle du physicien qui, pour calculer la trajectoire d'un objet dans le vide, ignore sa couleur ou sa composition chimique pour se concentrer uniquement sur ce qui compte, en l'occurrence sa masse et la gravité ; ou à celle du cartographe, qui modélise le paysage sans se préoccuper des brins d'herbe ou des grains de sable. C'est de cette manière qu'un réseau de neurones peut être implémenté dans une machine. Mais avant de modéliser le réseau, il faut modéliser son constituant : le neurone.

C'est à Warren S. McCulloch, neuroscientifique, et Walter Pitts, jeune logicien de génie, que l'on doit cette modélisation, présentée dans un article fondateur publié en 1943. Il ne s'agissait pas pour eux de produire une machine intelligente, mais de mieux comprendre le cerveau, quoique les possibles conséquences technologiques ne leur aient pas échappé, et qu'ils aient fait explicitement le lien avec l'informatique théorique naissante (ils citent notamment les travaux d'Alan Turing). En substance, McCulloch et Pitts modélisent les neurones comme une fonction mathématique très simple, une combinaison linéaire des signaux en entrée suivie d'une fonction d'activation (c'est en fait un peu plus compliqué dans leur article, mais restons simples).

Pas de panique : ce sont des mathématiques de collège ! Mais quelques rappels seront sans doute bienvenus.

Une combinaison linéaire, ce n'est rien de plus qu'une série de multiplications et d'additions. On prend des nombres en entrée (par exemple 1, 2 et 3), on les multiplie par d'autres nombres (par exemple 4, 5 et 6), on calcule la somme de tout ça (ce qui avec ces exemples donnerait $1 \times 4 + 2 \times 5 + 3 \times 6 = 32$), et on ajoute un dernier nombre à cette somme (par exemple -10 , ce qui donnerait $1 \times 4 + 2 \times 5 + 3 \times 6 - 10 = 22$). Une fonction d'activation, ça peut être simplement une instruction déterminant si un nombre en entrée doit être ou non transformé en 0. La fonction d'activation détermine donc si le neurone va renvoyer un nombre en sortie, et donc un signal (on parle alors de neurone « activé ») ou rien du tout (neurone inactif ou « inhibé »). Très souvent, cette fonction est un simple seuil : si le nombre en entrée, qui dans le cas du neurone sera le résultat de la combinaison linéaire (ici 22), est supérieur à 0, alors rien ne change en sortie, sinon on obtient 0. Cette fonction d'activation particulière s'appelle « ReLU » (« Rectified Linear Unit »). Il en existe d'autres un peu plus complexes, mais c'est toujours la même idée.

Les nombres en entrée (ici 1, 2 et 3) correspondent aux signaux électrochimiques que reçoivent les neurones biologiques via les synapses, le possible 22 en sortie correspond au signal émis par le neurone lorsqu'il est « activé » (le neurone est activé lorsque le seuil est franchi), les nombres qui viennent multiplier les signaux d'entrée (ici 4, 5 et 6) correspondent à la « force » de la connexion synaptique, ce que l'on appelle les « poids ». Plus ce poids est élevé, plus le signal d'entrée a d'influence sur l'activation du neurone. Enfin, le nombre qui vient s'ajouter à la somme des produits des entrées et des poids, ici -10 , correspond à l'opposé du seuil d'activation (qui dans cet exemple serait 10). Ce nombre s'appelle le « biais ».

Souvent, ces neurones formels sont représentés de la manière suivante :

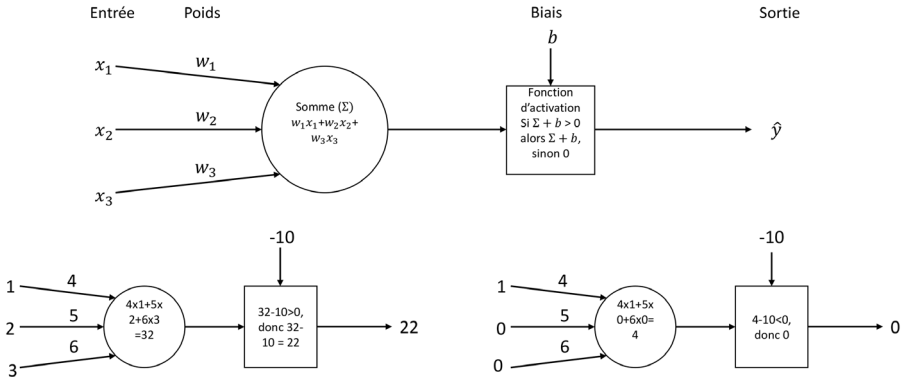


Figure 3. Schéma d'un neurone formel, avec deux exemples numériques illustrant un cas activé et un cas inhibé.

Mathématiquement, ces neurones peuvent être représentés par une fonction mathématique $\hat{y} = \varphi \left(\sum_i w_i \cdot x_i + b \right)$, φ étant la fonction d'activation, $\hat{y}$ le signal de sortie, x_i les signaux d'entrée (dans l'exemple ci-dessus, nous avons $x_1 = 1$, $x_2 = 2$ et $x_3 = 3$), w_i les poids (ici 4, 5 et 6), b le biais (ici $b = -10$). Les poids et le biais constituent ensemble les « paramètres » du système. Ici, nous avons quatre paramètres.

Un réseau de neurones est simplement l'assemblage de plusieurs neurones. Les sorties de certains neurones deviennent alors les entrées d'autres neurones. Ils sont généralement (pas toujours) organisés en « couches ». La couche qui reçoit les données à traiter est la « couche d'entrée », la couche qui produit le ou les résultats finaux et la « couche de sortie », et toutes les possibles couches intermédiaires sont appelées « couches cachées ». Les neurones de la couche d'entrée ne reçoivent qu'une seule entrée chacun et n'ont pas de fonction d'activation : ils servent surtout à injecter les données à traiter dans le reste du réseau. Tous les autres neurones reçoivent généralement leurs entrées des neurones de la couche précédente et sont dotés d'une fonction d'activation. La figure ci-dessous donne l'exemple d'un réseau avec quatre couches, trois entrées (parfois comptées parmi les « neurones »), cinq neurones dans

les deux couches cachées, et deux dans la couche de sortie. Ce réseau est « dense », car chaque neurone d'une couche est relié à chaque neurone de la couche suivante.

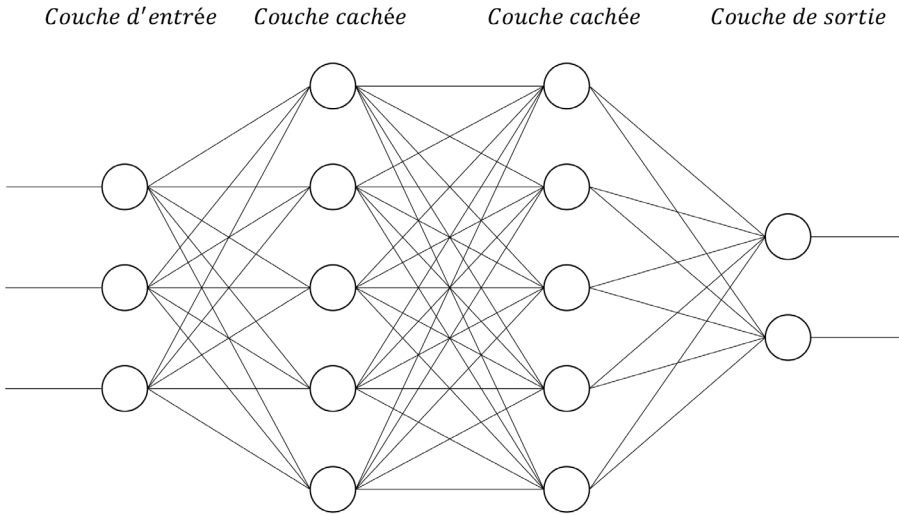


Figure 4. Perceptron multicouche avec deux couches cachées.

Ce réseau est en fait un exemple d'un type très important et très connu de réseau de neurones, déjà mentionné : c'est un perceptron multicouche (MLP). Par définition un MLP est dense, comporte au moins une couche cachée, et chaque neurone est lié aux neurones de la couche suivante *et seulement de la couche suivante* : pas à ceux de la couche d'après, pas à ceux de la couche d'avant, pas à lui-même ou à un neurone de la même couche. Lorsque le signal va toujours vers l'avant et ne retourne jamais en arrière, ce qui est donc le cas des MLP, on parle de réseaux de neurones « à action directe » (en anglais : « feedforward neural network »).

Lorsque le réseau de neurones ne comporte aucune couche cachée, et qu'il n'y a pas de fonction d'activation en sortie (ou une fonction linéaire $f(x) = x$, ce qui revient au même) on parle de « couche linéaire ». Une couche linéaire n'est rien de plus, mathématiquement parlant, qu'une combinaison linéaire. Nous allons voir que ces couches linéaires sont très importantes pour les LLM.