Intelligence artificielle

Melanie Mitchell

Intelligence artificielle

Triomphes et Déceptions

Postface de
Douglas Hofstadter

Traduit de l'anglais (États-Unis)
par Christian Jeanmougin

DUNOD

L'édition originale de cet ouvrage a été publiée en anglais sous le titre
Artificial Intelligence: A Guide for Thinking Human.

Avec la collaboration de Robert French

L'Éditeur remercie Patrick Géhant
pour l'aide apportée à la finalisation de la traduction.

Direction artistique : Nicolas Wiel
Image de couverture : © Jackie Niam / Adobe Stock.

Ouvrage publié avec le concours du

CNL CENTRE NATIONAL DU LIVRE

11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-081372-8

À mes parents, qui m'ont appris à penser, et bien plus encore.

Prologue

Terrifié

Si l'intelligence des ordinateurs semble croître à un rythme effrayant, il reste néanmoins une chose que ces machines ne savent pas encore faire : percevoir l'ironie. C'est ce que j'ai pensé il y a quelques années, quand, me rendant à une réunion sur l'intelligence artificielle (IA), je me suis perdue dans la capitale de la recherche et découverte – le Googleplex, le siège social de Google, à Mountain View, en Californie. Le comble, c'est que j'étais perdue à l'intérieur du bâtiment de Google Maps !

Le bâtiment lui-même avait été facile à trouver. Une voiture de Google Street View était garée près de l'entrée, portant sur son toit un imposant appendice surmonté d'une caméra en forme de ballon de foot rouge et noir. Mais une fois à l'intérieur, munie du célèbre badge « Visiteur » donné par la sécurité, j'ai erré, embarrassée, à travers une multitude de postes de travail occupés par des employés Google, écouteurs sur les oreilles et tapant frénétiquement sur des claviers d'ordinateurs Apple. Après une exploration aléatoire (sans plan), j'ai finalement trouvé la salle affectée à la réunion et rejoint les participants déjà arrivés.

Cette réunion, en mai 2014, avait été organisée par Blaise Agüera y Arcas, un jeune informaticien qui avait récemment quitté un poste important chez Microsoft pour aider Google à construire une intelligence artificielle. À sa création, en 1998, Google proposait un « produit » : un site web qui utilisait une méthode nouvelle, extraordinairement efficace, pour explorer le Web. Au fil des ans, Google

est devenu la plus importante compagnie technologique du monde et offre aujourd'hui un vaste éventail de produits et de services, tels Gmail, Google Docs, Google Translate, YouTube, Android, que vous utilisez peut-être chaque jour, et d'autres dont vous n'avez probablement jamais entendu parler.

Les fondateurs de Google, Larry Page et Sergey Brin, cherchent depuis longtemps à créer de l'intelligence artificielle dans les ordinateurs, et cet objectif est devenu une priorité majeure de Google. Durant la dernière décennie, cette entreprise a recruté de nombreux experts en IA, en particulier Ray Kurzweil, inventeur bien connu et futurologue controversé, qui affirme l'existence d'une Singularité technologique, d'un futur proche où les ordinateurs seront plus intelligents que les humains. Google a recruté Kurzweil pour l'aider à concrétiser cette vision. En 2011, Google créa un groupe de recherche interne en IA appelé Google Brain ; depuis, l'entreprise a également acquis un nombre impressionnant de startups en IA aux noms également optimistes tels que Applied Semantics, DeepMind, et Vision Factory.

Bref, Google n'est plus un simple portail de recherche – loin de là. Il se transforme rapidement en une entreprise d'application de l'IA. L'IA est la colle qui unifie les divers produits, services et recherches fondamentales proposés par Google et sa maison mère, Alphabet. L'ambition ultime de l'entreprise est résumée dans la formulation originelle de la mission de son groupe DeepMind : « Expliquer l'intelligence et l'utiliser pour expliquer tout le reste. »[1]

L'IA et *GEB*

J'étais passablement excitée à l'idée d'assister à une réunion sur l'IA chez Google. Depuis mes études doctorales dans les années 1980, je travaillais sur divers aspects de l'IA et j'étais énormément impressionnée par ce que Google avait accompli. Je pensais en outre avoir quelques idées intéressantes. Mais je dois admettre que ce jour-là, je n'étais qu'une simple spectatrice. Cette réunion devait permettre à quelques chercheurs en IA de chez Google, triés sur le volet, de dialoguer avec Douglas Hofstadter, légende vivante de l'IA et auteur d'un célèbre livre énigmatiquement

intitulé *Gödel, Escher, Bach : Les Brins d'une Guirlande Éternelle*, ou plus succinctement, *GEB*. Si vous êtes informaticien ou fan d'ordinateurs, vous avez probablement entendu parler de ce livre, à moins que vous ne l'ayez lu, ou essayé de le lire.

Écrit dans les années 1970, *GEB* fut une émanation des multiples passions intellectuelles d'Hofstadter – les mathématiques, l'art, la musique, le langage, l'humour, les jeux de mots, tout cela mobilisé pour se demander comment l'intelligence, la conscience et le sentiment de conscience de soi, dont chaque être humain a une expérience si profonde, peuvent émerger d'un substrat non intelligent, non conscient, composé de cellules biologiques. Il se demandait également comment les ordinateurs pourraient finalement acquérir une intelligence et une conscience de soi. C'est un livre unique ; je n'en connais aucun autre qui lui soit comparable, même de loin. Bien qu'il ne soit pas facile à lire, il est pourtant devenu un best-seller et a obtenu le prix Pulitzer et le National Book Award. Incontestablement, *GEB* est plus que tout autre livre celui qui a le plus incité de jeunes gens à s'intéresser à l'IA. Je fus l'une d'entre eux.

Au début des années 1980, après avoir obtenu ma licence de maths, j'ai vécu à New York, enseignant les maths dans une école privée, malheureuse, me demandant ce que je voulais réellement faire dans la vie. Je découvris *GEB* après en avoir lu une critique dithyrambique dans *Scientific American*. Je sortis immédiatement l'acheter. Je le dévorai en quelques semaines, de plus en plus convaincue non seulement que je voulais devenir chercheuse en IA, mais surtout que je voulais travailler avec Douglas Hofstadter. Je n'avais jamais été autant passionnée par un livre ni persuadée de faire le bon choix pour ma carrière.

À l'époque, Hofstadter enseignait l'informatique à l'Université d'Indiana, et mon plan chimérique consistait à m'y inscrire en doctorat d'informatique, puis à me rendre sur place et convaincre Hofstadter de m'accepter comme étudiante. Il y avait cependant un petit problème : je n'avais jamais suivi de cours d'informatique. J'avais grandi parmi des ordinateurs ; mon père était ingénieur matériel dans une start-up technologique des années 1960 et avait construit un gros ordinateur dans le coin repos familial. Cette machine, une Scientific Data Systems Sigma 2, de la taille d'un réfrigérateur, portait un magnet proclamant

« Je prie en FORTRAN », et enfant, j'étais pratiquement convaincue qu'il le faisait, discrètement, la nuit, pendant que la famille dormait. Dans les années 1960 et 1970, j'ai appris quelques rudiments de langages de programmation populaires à l'époque – le FORTRAN, puis le BASIC, puis le Pascal –, mais je ne savais pratiquement rien des techniques de programmation proprement dites, et moins que rien de tout ce que doit en outre savoir une future diplômée en informatique.

Pour lancer mon projet, je quittai mon poste d'enseignante à la fin de l'année scolaire, déménageai à Boston, et commençai à suivre des cours d'introduction à l'informatique pour préparer ma nouvelle carrière. Quelques mois plus tard, j'étais sur le campus du Massachusetts Institute of Technology, attendant le début d'un cours, lorsque je vis une affiche annonçant une conférence de Douglas Hofstadter deux jours plus tard sur ce même campus. Je n'en croyais pas mes yeux ; la chance me souriait. J'assistai à la conférence puis, après une longue attente parmi la foule des admirateurs, je parvins à parler à Hofstadter. Il s'avéra qu'il était en année sabbatique au MIT, au terme de laquelle il déménagerait d'Indiana à l'Université du Michigan, à Ann Arbor.

Pour faire bref, après des demandes répétées de ma part, je le persuadai de m'engager comme assistante de recherche, d'abord pour un été, puis pour les six années suivantes en tant qu'étudiante de troisième cycle, au terme desquelles je passai un doctorat en informatique à l'Université du Michigan. Hofstadter et moi sommes restés en étroits contacts au fil des ans et avons eu de nombreuses discussions sur l'IA. Connaissant mon intérêt pour les recherches de Google en IA, il m'avait gentiment invitée à la réunion organisée par Google.

Le jeu d'échecs : première apparition du doute

Le groupe dans la salle de conférences pas facile à trouver se composait d'une vingtaine d'ingénieurs Google (plus Douglas Hofstadter et moi-même), tous membres de diverses équipes IA de Google. La rencontre débuta par le traditionnel tour de salle au cours duquel les participants se présentent rapidement. Plusieurs d'entre eux précisèrent que leur propre carrière avait été stimulée par la lecture de *GEB* durant leur jeunesse. Tous étaient excités et curieux d'entendre ce que le légendaire

Hofstadter avait à dire sur l'IA. Hofstadter se leva alors pour parler : « J'ai quelques remarques à faire sur la recherche en IA en général, et ici chez Google en particulier. » Sa voix se fit passionnée. « Je suis terrifié. Terrifié. »

Hofstadter poursuivit[2]. Dans les années 1970, expliqua-t-il, quand il commença à s'y intéresser, l'IA était une perspective exaltante mais semblait si loin de se concrétiser qu'il n'y avait aucun « danger à l'horizon, aucun sentiment qu'elle *survienne* réellement. » La création de machines douées d'intelligence humaine était une profonde aventure intellectuelle, un projet de recherche à long terme, dont la réalisation, disait-on, se situait à au moins « cent prix Nobel de nous »[3]. Hofstadter croyait que l'IA était possible en principe : « L'"ennemi", c'était des gens comme John Searle, Hubert Dreyfus et d'autres sceptiques, qui disaient qu'elle était impossible. Ils ne comprenaient pas qu'un cerveau est un morceau de matière qui obéit à des lois physiques et que l'ordinateur peut tout simuler… le niveau des neurones, des neurotransmetteurs, et cætera. En principe, c'est faisable. » En fait, les idées d'Hofstadter sur la simulation de l'intelligence à divers niveaux – des neurones jusqu'à la conscience – sont discutées en détail dans *GEB* et furent au cœur de ses propres recherches durant des décennies. Mais en pratique, jusqu'à récemment, Hofstadter pensait qu'une IA générale de « niveau humain » n'avait aucune chance de voir le jour durant sa vie (voire celle de ses enfants), de sorte qu'il ne se souciait pas trop de cet aspect des choses.

Vers la fin de *GEB*, Hofstadter a listé « Dix questions et réflexions » concernant l'intelligence artificielle. L'une de ces questions demande : « Existera-t-il des programmes d'échecs capables de battre n'importe quel adversaire humain ? » La réponse d'Hofstadter était « non ». « Il existera peut-être des programmes battant tout le monde aux échecs, mais ce ne seront pas exclusivement des programmes d'échecs. Il s'agira de programmes d'intelligence *générale* »[4].

Lors de la réunion Google, en 2014, Hofstadter a reconnu qu'il s'était « totalement trompé ». Le rapide perfectionnement des programmes d'échecs, dans les années 1980 et 1990, avait commencé à instiller le doute dans sa vision des progrès de l'IA à court terme. Bien qu'Herbert Simon, l'un des pionniers de l'IA, ait prédit

en 1957 qu'un programme d'échecs serait champion du monde « d'ici dix ans », au milieu des années 1970, alors que Hofstadter rédigeait *GEB*, les meilleurs programmes d'échecs avaient seulement atteint le niveau d'un bon amateur. Eliot Hearst, champion d'échecs, professeur de psychologie et ami d'Hofstadter, avait abondamment écrit sur les différences entre grands joueurs d'échecs et programmes d'échecs informatiques. Les expériences montraient que pour décider d'un coup, les grands joueurs utilisent la reconnaissance rapide des configurations sur l'échiquier plutôt que le recours systématique à la force d'anticipation brute qu'utilisent tous les programmes d'échecs. Durant une partie, les meilleurs joueurs humains voient dans une configuration de pièces un « type de position » exigeant un certain « type de stratégie ». Autrement dit, ces joueurs peuvent rapidement identifier dans des configurations et stratégies particulières des instances de concepts de niveaux supérieurs. Hearst affirmait que tant que les ordinateurs ne posséderaient pas cette aptitude générale à percevoir des configurations et à reconnaître des concepts abstraits, les programmes d'échecs n'atteindraient jamais le niveau des meilleurs joueurs humains. Hofstadter était d'accord avec Hearst.

Toutefois, dans les années 1980 et 1990, les programmes d'échecs devinrent bien plus performants, grâce principalement à un net accroissement de la vitesse des ordinateurs. Les meilleurs programmes jouaient encore de manière très inhumaine : en recourant toujours à la force d'anticipation brute pour décider du coup suivant. Au milieu des années 1990, la machine Deep Blue, d'IBM, dotée d'un matériel conçu spécifiquement pour le jeu d'échecs, avait atteint le niveau de grand maître, et en 1997, elle battit le champion du monde en titre, Garry Kasparov, lors d'une rencontre en six parties. La maîtrise aux échecs, qui autrefois semblait être le summum de l'intelligence humaine, avait succombé devant la recherche par la force brute.

La musique, bastion de l'humanité

Bien que la victoire de Deep Blue, marque d'une ascension des machines intelligentes, ait provoqué une vague de consternation et d'inquiétude dans la presse, la « vraie » IA semblait encore très lointaine. Deep Blue

savait jouer aux échecs, mais elle ne savait rien faire d'autre. Hofstadter s'était trompé sur les échecs, mais n'avait pas changé de point de vue sur les autres questions posées dans *GEB*, en particulier sur la première de sa liste :

> QUESTION : un ordinateur pourra-t-il écrire de la belle musique ?
> RÉFLEXION : oui, mais pas avant longtemps.

Hofstadter poursuivit :

> La musique est un langage d'émotions, et tant que les programmes n'éprouveront pas d'émotions aussi complexes que les nôtres, il est impossible qu'un programme écrive quoi que ce soit de beau. Des programmes peuvent écrire des « contrefaçons », de pâles imitations de la syntaxe de la musique composée par d'autres, mais en dépit de ce que l'on pourrait *a priori* penser, les règles syntaxiques ne font pas l'essence de la musique. […] Penser […] que nous pourrons bientôt commander à une « boîte à musique » de bureau préprogrammée, fabriquée en série, achetée par correspondance, et bon marché, de sortir de ses circuits stériles des morceaux que Chopin ou Bach auraient pu écrire s'ils avaient vécu plus longtemps, c'est commettre une erreur d'appréciation grotesque et éhontée sur la profondeur de l'esprit humain[5].

Pour Hofstadter, cette réflexion constituait « l'une des parties les plus importantes de *GEB* – j'aurais parié ma vie sur elle. »

Au milieu des années 1990, sa confiance en son évaluation de l'IA fut de nouveau ébranlée, cette fois très profondément, lorsqu'il tomba sur un programme écrit par un musicien nommé David Cope. Ce programme s'appelait « Experiments in Musical Intelligence » (EMI, Expériences en intelligence musicale). Cope, compositeur et professeur de musique, avait initialement développé EMI pour l'aider dans son travail de composition en créant automatiquement des morceaux dans son propre style. EMI est toutefois devenu célèbre pour avoir créé des morceaux dans le style de compositeurs classiques tels que Bach et Chopin. Il compose en suivant un vaste ensemble de règles, conçues par Cope pour définir explicitement une syntaxe générale de composition. Appliquées à un vaste échantillonnage de l'œuvre d'un compositeur, elles visent à produire une œuvre nouvelle « dans le style » de ce compositeur.

Lors de la réunion Google, Hofstadter parla avec une grande émotion de ses rencontres avec EMI :

> Je me suis mis à mon piano et j'ai joué l'une des mazurkas écrites par EMI « dans le style de Chopin ». Ça ne sonnait pas exactement comme du Chopin, mais ça sonnait suffisamment comme du Chopin, et comme de la musique cohérente, pour que j'en sois *profondément* troublé.
> Depuis mon enfance, la musique me transporte et m'émeut jusqu'au plus profond de moi-même. Et chacune de mes œuvres préférées me semble être un message directement envoyé du cœur émotionnel de l'être humain qui l'a composée. Elle semble m'ouvrir la partie la plus intime de son âme. Et il semble n'y avoir *rien* de plus humain dans le monde que l'expression de la musique. Rien. L'idée que la plus superficielle des manipulations de formes puisse donner des choses semblant venir du cœur d'un être humain est très, très troublante. Elle me sidérait totalement.

Hofstadter parla ensuite d'une conférence qu'il avait donnée à la prestigieuse Eastman School of Music, à Rochester, dans l'État de New York. Après avoir décrit EMI, Hofstadter demanda à son public – qui comprenait plusieurs professeurs de théorie musicale et de composition – de deviner lequel des deux morceaux joués devant eux par un pianiste était une mazurka (peu connue) de Chopin et lequel était une composition de EMI. Comme l'expliqua par la suite un membre du public, « la première mazurka avait de la grâce et du charme, mais n'avait pas l'inventivité et la grande fluidité typiques de Chopin […]. La seconde, par contre, avec son lyrisme mélodique, ses gracieuses et amples modulations chromatiques, et sa forme naturellement équilibrée, était incontestablement du Chopin. »[6] Bon nombre d'enseignants partagèrent ce point de vue et, à la stupéfaction d'Hofstadter, votèrent EMI pour le premier morceau et « véritable Chopin » pour le second. Les bonnes réponses étaient exactement l'inverse.

Dans la salle de conférences de Google, Hofstadter marqua une pause et nous scruta du regard. Personne ne disait mot. Finalement, il reprit : « J'étais terrifié par EMI. Terrifié. Je le haïssais et me sentais

extrêmement menacé par lui. Il menaçait de détruire ce que je chérissais le plus dans l'humanité. Je pense que ce programme était la quintessence même des craintes que suscite en moi l'intelligence artificielle. »

Google et la Singularité

Hofstadter parla alors de sa profonde ambivalence à l'égard de ce que Google elle-même tentait d'accomplir en IA – notamment, la voiture autonome, la reconnaissance de la parole, la compréhension du langage naturel, la traduction entre diverses langues, la création artistique par ordinateur, la composition musicale, etc. Les inquiétudes d'Hofstadter se sont accentuées lorsque Google a accueilli Ray Kurzweil et sa vision de la Singularité, selon laquelle l'IA, utilisant sa capacité à se perfectionner et apprendre toute seule, atteindra rapidement, puis dépassera, l'intelligence humaine. Il semblait que Google faisait tout ce qu'elle pouvait pour concrétiser cette vision le plus vite possible. Si Hofstadter doutait fortement de l'hypothèse de la Singularité, il reconnaissait néanmoins être troublé par les prédictions de Kurzweil. « J'étais terrifié par les scénarios. Je restais très sceptique, mais en même temps, je me disais peut-être qu'ils ont raison, même s'ils se trompent sur la date d'arrivée de la Singularité. Nous serons totalement pris au dépourvu. Nous penserons qu'il ne se passe rien et tout d'un coup les ordinateurs seront plus intelligents que nous. »

Si cela se produit réellement, « nous serons supplantés. Nous serons des reliques. Nous serons complètement largués par les machines. »

« Cela arrivera peut-être, mais je ne veux pas que cela arrive *bientôt*. Je ne veux pas que mes enfants soient complètement dépassés par les ordinateurs. »

Hofstadter conclut son propos par une référence directe aux ingénieurs Google présents dans la salle, suspendus à ses paroles : « Je trouve très angoissant, très troublant, très triste, atroce, effrayant, bizarre, déroutant, incompréhensible que des gens se précipitent, aveuglément et de manière complètement délirante, pour créer ces choses. »

D'où vient la terreur d'Hofstadter ?

J'ai regardé autour de moi. L'auditoire était décontenancé, embarrassé même. Pour ces chercheurs en IA, travaillant chez Google, rien de tout cela n'était le moins du monde terrifiant. En fait, c'était une vieille histoire. Lorsque Deep Blue battit Kasparov, lorsque EMI commença à composer des mazurkas dans le style de Chopin et que Kurzweil écrivit son premier livre sur la Singularité, nombre de ces ingénieurs étaient au lycée, se délectant probablement de la lecture de *GEB* même si ses pronostics sur l'IA étaient un peu dépassés. La raison même pour laquelle ils travaillaient chez Google était justement pour faire advenir l'IA – non dans une centaine d'années, mais aujourd'hui, le plus tôt possible. Ils ne comprenaient pas pourquoi Hofstadter était si stressé.

Les chercheurs en IA ont l'habitude d'entendre les inquiétudes des gens extérieurs à leur discipline, sans doute influencés par les nombreux films de science-fiction montrant des machines super-intelligentes se retourner contre leurs maîtres. Il leur arrive également d'entendre des gens craindre qu'une IA de plus en plus sophistiquée ne remplace les humains dans certaines professions, que son application aux mégadonnées ne porte atteinte à la vie privée et n'ouvre la voie à une discrimination subtile, et que des systèmes d'IA mal maîtrisés, mais néanmoins autorisés à prendre des décisions autonomes, ne sèment le chaos.

La terreur d'Hofstadter était une réaction à une chose entièrement différente. Elle ne concernait pas une IA devenant trop intelligente, trop invasive, trop malveillante ou même trop utile. Non, ce qui le terrifiait était que l'intelligence, la créativité, les émotions, voire la conscience elle-même, soient trop *faciles* à engendrer – que ce qu'il appréciait le plus dans l'humanité s'avère finalement n'être rien d'autre qu'un « sac à malices », qu'un ensemble superficiel d'algorithmes bruts qui parviendraient à expliquer l'esprit humain.

Et pourtant, dans *GEB*, Hofstadter le dit sans ambiguïté : l'esprit et toutes ses caractéristiques émergent uniquement du substrat physique du cerveau et du reste du corps, ainsi que de l'interaction du corps avec le monde physique. Autrement dit, il n'y a rien d'immatériel ou d'incorporel dissimulé dans cette émergence. Le problème

qui le travaille est réellement celui de la complexité. Il craint que l'IA ne nous révèle que les qualités humaines que nous apprécions le plus soient désespérément simples à mécaniser. Comme il me l'a expliqué après la réunion en évoquant Chopin, Bach et d'autres parangons d'humanité, « Si de tels esprits, d'une subtilité, d'une complexité et d'une profondeur émotionnelle infinies, pouvaient être banalisés par une petite puce électronique, cela détruirait mon sentiment profond de ce que nous sommes, de notre humanité. »

Je suis perplexe

Après les remarques d'Hofstadter, il y eut une brève discussion au cours de laquelle les ingénieurs Google, déconcertés, poussèrent Hofstadter à préciser ses craintes concernant l'IA et Google en particulier. Mais une barrière de communication demeura. La réunion se poursuivit, avec présentations de projets, discussions de groupe, pauses-café, la routine habituelle, mais rien ne porta sur les commentaires d'Hofstadter. Vers la fin de la réunion, Hofstadter demanda à l'auditoire comment il voyait l'avenir à court terme de l'IA. Plusieurs ingénieurs prédirent qu'une IA de niveau humain général émergerait probablement d'ici les trente prochaines années, en grande partie grâce aux progrès accomplis par Google en « réseaux de neurones artificiels profonds ».

J'ai quitté la réunion ne sachant plus très bien que penser. Je savais qu'Hofstadter avait été troublé par certains écrits de Kurzweil sur la Singularité, mais je n'avais jamais pleinement compris la profondeur de son émotion et de son anxiété. Je savais également que Google faisait de gros efforts de recherche en IA, mais j'étais étonnée par l'optimisme de plusieurs de ses ingénieurs à l'égard de la rapidité avec laquelle ils estimaient que l'IA atteindrait un niveau « humain » général. Jusque-là, je pensais que l'IA avait extrêmement progressé dans certains domaines limités, mais qu'elle était encore loin d'avoir l'intelligence générale et diversifiée des humains, et ne l'atteindrait pas dans un siècle, et encore moins dans trente ans. Et je pensais que les gens qui croyaient le contraire sous-estimaient grandement la complexité de l'intelligence humaine. J'avais lu des livres de Kurzweil et les avais trouvés en grande partie ridicules. Mais tous les commentaires que j'ai entendus lors de

la réunion, provenant de personnes que je respectais et que j'admirais, m'obligèrent à examiner de manière critique mes propres points de vue. Si je supposais que ces chercheurs en IA sous-estimaient les êtres humains, peut-être de mon côté avais-je sous-estimé le pouvoir et l'avenir de l'actuelle IA ?

Dans les mois qui suivirent, je fis davantage attention aux débats entourant ces questions. J'ai commencé à remarquer la multitude d'articles, de blogs et de livres entiers rédigés par d'éminentes personnes nous disant soudainement que nous devrions nous inquiéter, tout de suite, des dangers liés à une IA « surhumaine », dépassant les capacités intellectuelles humaines. En 2014, le physicien Stephen Hawking déclara : « Le développement de la vraie intelligence artificielle pourrait entraîner la fin de l'espèce humaine. »[7] Cette même année, l'entrepreneur Elon Musk, fondateur des entreprises Tesla et SpaceX, affirma que l'intelligence artificielle est probablement « notre plus grande menace existentielle » et qu'« avec l'intelligence artificielle, nous invoquons le démon. »[8] Le cofondateur de Microsoft, Bill Gates, approuva : « Je suis d'accord sur ce point avec Elon Musk et quelques autres, et je ne comprends pas pourquoi des gens ne se sentent pas concernés. »[9] Le livre du philosophe Nick Bostrom, *Superintelligence*, sur les dangers potentiels des machines qui deviennent plus intelligentes que les humains, fut un best-seller surprise, malgré l'aridité et la lourdeur de son style

D'autres éminents penseurs tentaient de contrer cet alarmisme. Oui, disaient-ils, nous devrions nous assurer que les programmes d'IA sont sûrs et ne risquent pas de nuire aux humains, mais toute possibilité d'une IA surhumaine à court terme est grandement exagérée. Selon l'entrepreneur et activiste Mitchell Kapor, « l'intelligence humaine est un phénomène merveilleux, subtil et mal compris. On ne risque pas de la dupliquer avant longtemps. »[10] Le roboticien (et ancien directeur du Laboratoire IA du MIT) Rodney Brooks approuve : nous « surestimons nettement les capacités des machines – celles d'aujourd'hui et celles des prochaines décennies » a-t-il écrit[11]. Le psychologue et chercheur en IA Gary Marcus est allé jusqu'à affirmer que la recherche d'une « IA forte » – c'est-à-dire d'une IA de niveau humain *général* – « n'a connu pratiquement aucun progrès. »[12]

Je pourrais poursuivre indéfiniment ce duel de citations. En bref, ce que j'ai trouvé, c'est que la discipline qu'on appelle « Intelligence Artificielle » est très controversée à l'heure actuelle. Soit on y a accompli d'énormes progrès, soit pratiquement aucun. Soit nous sommes à deux pas de la « vraie » IA, soit nous en sommes éloignés de plusieurs siècles. Soit l'IA résoudra nos problèmes, soit elle nous mettra tous au chômage, détruira l'espèce humaine, ou dévalorisera notre humanité. Elle est soit une noble quête, soit une « invocation du démon ».

De quoi parle ce livre

Ce livre est né de mon désir de comprendre la situation dans laquelle se trouve réellement l'intelligence artificielle – ce que peuvent faire actuellement les ordinateurs, et ce que nous pouvons en attendre dans les prochaines décennies. Les commentaires provocateurs d'Hofstadter lors de la réunion chez Google ont déclenché en moi une prise de conscience, tout comme les réponses confiantes des chercheurs Google sur l'avenir à court terme de l'IA. Dans les chapitres qui suivent, je tente de voir où en est l'intelligence artificielle et de clarifier ses objectifs disparates – et parfois contradictoires. Ce faisant, j'examine le fonctionnement réel de certains des plus importants systèmes IA et recherche ce qui fait leur efficacité et leurs limitations. Je regarde dans quelle mesure les ordinateurs peuvent aujourd'hui faire des choses qui exigent selon nous de hauts niveaux d'intelligence – battre des humains aux jeux les plus intellectuellement exigeants, traduire d'une langue dans une autre, répondre à des questions complexes, conduire des véhicules en terrain difficile. J'examine également leur comportement dans des situations qui nous semblent aller de soi, dans les tâches de la vie quotidienne que nous accomplissons sans y penser, telles que la reconnaissance de visages et d'objets sur des images, la compréhension du langage parlé et du texte écrit, et l'utilisation du bon sens le plus élémentaire.

Je m'efforce également de donner un sens aux questions plus vastes qui alimentent les débats sur l'IA depuis sa création. Qu'entendons-nous effectivement par intelligence « humaine générale », voire « surhumaine » ? L'actuelle IA est-elle proche de ce niveau, voire sur une trajectoire qui pourrait y conduire ? Quels sont les dangers ? Quels aspects de notre

intelligence nous sont les plus chers, et dans quelle mesure l'IA de niveau humain mettrait en question notre perception de notre propre humanité ? Pour parler comme Douglas Hofstadter, jusqu'à quel point devrions-nous être « terrifiés » ?

Ce livre n'est pas une étude générale ou une histoire de l'intelligence artificielle. Il est plutôt une exploration en profondeur de quelques méthodes utilisées en IA qui probablement influent ou influeront bientôt sur votre vie, ainsi que des efforts de l'IA qui vont peut-être le plus loin dans la remise en question de notre sentiment d'unicité de l'espèce humaine. Mon but est de vous amener à partager cette exploration et, comme moi, de repartir avec un sentiment plus clair de ce que cette discipline a accompli et du chemin qui reste à parcourir avant que nos machines puissent se prétendre dotées d'une humanité propre.

Première partie

Le contexte

1

Les racines de l'intelligence artificielle

Deux mois et dix hommes à Dartmouth

Le rêve de créer une machine intelligente – aussi intelligente, voire plus intelligente, que les humains – est vieux de plusieurs siècles, mais a intégré la science moderne avec l'arrivée des ordinateurs. En fait, les idées qui ont conduit aux premiers ordinateurs programmables ont résulté de tentatives faites par des mathématiciens pour comprendre la pensée humaine – en particulier la logique – en tant que processus mécanique de « manipulation de symboles ». Les ordinateurs sont essentiellement des manipulateurs de symboles qui jonglent avec des combinaisons des symboles 0 et 1. Pour les pionniers de l'informatique tels que Alan Turing et John von Neumann, il y avait de grandes analogies entre les ordinateurs et le cerveau humain, et il leur semblait évident que l'on pouvait reproduire l'intelligence humaine dans des programmes informatiques.

La plupart des chercheurs en intelligence artificielle font remonter la fondation officielle de leur discipline à un petit atelier organisé en 1956 à Dartmouth College par un jeune mathématicien nommé John McCarthy.

En 1955, âgé de vingt-huit ans, McCarthy rejoignit la faculté de mathématiques de Dartmouth. Durant ses études de premier cycle, il avait acquis quelques rudiments de psychologie et de « théorie

des automates » – une toute nouvelle discipline qui allait devenir l'informatique –, et avait caressé l'idée de créer une machine pensante. Durant ses études de troisième cycle dans le département de mathématiques de l'Université de Princeton, McCarthy fit la connaissance d'un étudiant, Marvin Minsky, qui partageait sa fascination pour le potentiel des ordinateurs intelligents. Après ses études, McCarthy travailla quelque temps aux Laboratoires Bell et chez IBM, où il collabora avec Claude Shannon, l'inventeur de la théorie de l'information, et Nathaniel Rochester, l'un des pionniers du génie électrique. Une fois à Dartmouth, McCarthy persuada Minsky, Shannon et Rochester de l'aider à organiser « durant l'été 1956, un workshop de 2 mois, réunissant 10 hommes et traitant de l'intelligence artificielle. »[1] L'expression *intelligence artificielle* est due à McCarthy ; il voulait distinguer cette discipline des recherches connexes regroupées sous le nom de cybernétique[2]. Par la suite, McCarthy reconnut que personne n'aimait vraiment ce nom – après tout, le but était *l'authentique* intelligence, et non l'« artificielle » – mais « je devais lui donner un nom, alors je l'ai appelée "intelligence artificielle". »[3]

Les quatre organisateurs soumirent une proposition à la Fondation Rockefeller, accompagnée d'une demande de subvention pour leur workshop, qui reposait sur « la conjecture selon laquelle chaque aspect de l'apprentissage ou tout autre trait de l'intelligence peut en principe être décrit avec une précision telle que l'on peut construire une machine capable de le simuler. »[4] La proposition énumérait une série de sujets à discuter – le traitement du langage naturel, les réseaux neuronaux, l'apprentissage automatique, les concepts abstraits et le raisonnement, la créativité – qui définissent encore aujourd'hui la discipline.

Même si les ordinateurs les plus avancés en 1956 étaient environ un million de fois plus lents que les téléphones portables d'aujourd'hui, McCarthy et ses collègues étaient persuadés que l'IA était à portée de main : « Nous pensons que l'on peut accomplir un progrès significatif dans un ou plusieurs de ces problèmes pour peu qu'un groupe de scientifiques triés sur le volet y travaillent ensemble durant un été. »[5]

Des obstacles apparurent rapidement, que connaîtrait aujourd'hui tout organisateur de workshop. La Fondation Rockefeller n'accorda que la moitié de la subvention demandée. Et il s'avéra plus difficile que

ne le pensait McCarthy de persuader les participants de venir séjourner à Dartmouth, voire de s'entendre sur quoi que ce soit. Il y eut beaucoup de discussions intéressantes, mais peu de cohérence. Comme d'habitude avec ce genre de rencontres, « chacun avait une idée différente, un solide ego, et beaucoup d'enthousiasme pour son propre projet. »[6] L'été de l'IA à Dartmouth donna néanmoins quelques résultats très importants. Cette discipline reçut un nom, et l'on esquissa ses objectifs généraux. Ceux qui parmi ses pionniers allaient bientôt devenir les *big four* (les quatre grands) – McCarthy, Minsky, Allen Newell et Herbert Simon – se retrouvèrent pour planifier l'avenir. Et l'on ne sait pour quelle raison, tous quatre quittèrent la réunion débordants d'optimisme pour leur discipline. Au début des années 1960, McCarthy fonda le Laboratoire d'intelligence artificielle de l'Université de Stanford, dont « l'objectif était de construire en une décennie une machine totalement intelligente. »[7] À la même époque, le futur prix Nobel Herbert Simon prédit que « d'ici vingt ans, des machines seront capables de faire n'importe quel travail actuellement réalisable par un humain. »[8] Peu après, Marvin Minsky, fondateur du MIT AI Lab (Laboratoire d'IA du MIT), déclara que « d'ici une génération, […] les problèmes posés par la création d'une "intelligence artificielle" seront en grande partie résolus. »[9]

On définit, puis on va de l'avant

Aucun de ces événements annoncés ne s'est encore réalisé. Où en sommes-nous alors de la construction d'une « machine totalement intelligente » ? Une telle machine nous obligerait-elle à rétro-concevoir le cerveau humain dans toute sa complexité, ou y a-t-il un raccourci, un ensemble intelligent d'algorithmes encore inconnus, susceptible de produire ce que nous reconnaîtrions comme de « l'intelligence totale » ? Que signifie même « intelligence totale » ?

« Définissez les termes, vous dis-je, ou jamais nous ne nous entendrons. »[10] Cet avertissement de Voltaire est un défi pour quiconque parle d'intelligence artificielle, car sa notion centrale – l'intelligence – est encore extrêmement mal définie. Marvin Minsky lui-même est allé jusqu'à forger l'expression « mot-valise »[11] pour qualifier des termes

tels que *intelligence* et ses nombreux cousins, tels *pensée*, *cognition*, *conscience* et *émotion*, chacun étant comme une valise contenant un fouillis de différents sens. En arborant différents sens en fonction du contexte, *intelligence artificielle* hérite de ce problème.

La plupart des gens conviendraient que les humains sont intelligents et que les grains de poussière ne le sont pas. De même, nous considérons généralement que les humains sont plus intelligents que les vers de terre. Pour ce qui est de l'intelligence humaine, le QI se mesure sur une seule échelle, mais nous parlons aussi de différentes dimensions – émotionnelle, verbale, spatiale, logique, artistique, etc. – de l'intelligence. Ainsi, l'intelligence peut être binaire (une chose l'est ou ne l'est pas), continue (une chose est plus intelligente qu'une autre) ou multidimensionnelle (une personne peut avoir une grande intelligence verbale mais une faible intelligence émotionnelle). En fait, le mot *intelligence* est une valise hyper-bondée, avec sa fermeture éclair prête à sauter.

Pour le meilleur ou pour le pire, l'IA a grandement ignoré ces diverses distinctions et s'est plutôt concentrée sur deux types d'activités, l'un d'ordre scientifique, l'autre d'ordre pratique. Côté scientifique, les chercheurs en IA étudient les mécanismes de l'intelligence « naturelle » (c'est-à-dire biologique) en essayant de les programmer dans les ordinateurs. Côté pratique, les partisans de l'IA veulent simplement créer des programmes informatiques qui effectuent des tâches aussi bien ou mieux que les humains, et ne se soucient pas de savoir si ces programmes *pensent* réellement de la même manière que les humains. Quand on leur demande si leurs motivations sont d'ordre pratique ou scientifique, nombre de chercheurs en IA répondent en plaisantant que cela dépend de l'origine de leurs subventions.

Dans un récent rapport sur l'état actuel de l'IA, un comité d'éminents chercheurs a défini cette discipline comme « une branche de l'informatique qui étudie les propriétés de l'intelligence en synthétisant l'intelligence. »[12] Un peu circulaire, n'est-ce pas ? Mais ce même comité a également reconnu qu'il est difficile de définir cette discipline, et c'est peut-être une bonne chose : « L'absence de définition précise, universellement admise, de l'IA a probablement contribué au développement, à l'épanouissement et au progrès de cette discipline à un

rythme sans cesse croissant. »[13] En outre, remarque le comité, « les praticiens, chercheurs spécialistes et développeurs de l'IA sont plutôt guidés par un vague sens de l'orientation et par l'impératif "d'aller de l'avant" ».

Une anarchie de méthodes

Au workshop de Dartmouth, en 1956, les participants n'avaient pas une vision unanime de l'approche à adopter pour développer l'IA. Certains – généralement des mathématiciens – considéraient que la logique mathématique et le raisonnement déductif étaient le langage de la pensée rationnelle. D'autres étaient partisans de méthodes inductives dans lesquelles les programmes extraient des statistiques à partir des données et utilisent les probabilités pour gérer l'incertitude. D'autres encore croyaient fermement qu'il fallait s'inspirer de la biologie et de la psychologie pour créer des programmes calqués sur la structure du cerveau. Cela vous surprendra peut-être, mais les arguments avancés par les partisans de ces diverses approches n'ont pas changé depuis Dartmouth. Et chaque approche a généré sa propre panoplie de principes et de techniques, complétée par des conférences et des revues spécialisées, avec peu d'échanges entre les sous-spécialités. Une récente recension de la littérature sur l'IA a résumé ainsi la situation : « Comme nous ne comprenons pas suffisamment ce qu'est l'intelligence ou ne savons pas produire une IA de niveau général, plutôt que d'arrêter certaines voies de recherche, nous devrions, pour véritablement progresser, recourir à l'"anarchie de méthodes" que constitue l'IA. »[14]

Mais depuis les années 2010, une famille de méthodes en IA – collectivement appelées « apprentissage profond » (ou réseaux neuronaux profonds) – s'est élevée au-dessus de cette anarchie pour devenir le paradigme dominant au sein de l'IA. En fait, dans la plupart des médias populaires, l'expression *intelligence artificielle* elle-même en est venue à signifier « apprentissage profond ». Cette confusion est regrettable et il me faut la dissiper. L'IA est une discipline qui comprend une multitude d'approches visant à créer des machines douées d'intelligence. L'apprentissage profond n'est que l'une de ces approches – l'une des nombreuses méthodes utilisées en *apprentissage machine*,

sous-discipline de l'IA, dans laquelle les machines « apprennent » à partir de données ou de leurs propres « expériences ». Pour mieux comprendre ces diverses différences, il importe de comprendre les causes d'une scission philosophique survenue dans les premiers temps de la communauté des chercheurs en IA : la scission entre l'IA dite symbolique et l'IA sub-symbolique.

L'IA symbolique

Regardons d'abord l'IA *symbolique*. La connaissance d'un programme d'IA symbolique se compose de mots ou de phrases (les « symboles ») généralement compréhensibles par un humain, ainsi que de règles selon lesquelles ce programme combine et traite ces symboles afin d'effectuer la tâche qui lui est affectée.

Je vous donne un exemple. L'un des premiers programmes d'IA fut baptisé en toute confiance General Problem Solver[15], GPS en abrégé, et désigné ainsi car ces créateurs pensaient qu'il s'agissait d'un algorithme général de résolution de toute sorte de problèmes. (Désolée pour cette collision d'acronymes ; le General Problem Solver est antérieur au Global Positioning System.) Et en effet, ce GPS pouvait résoudre des problèmes comme celui « des missionnaires et des cannibales », que vous avez peut-être déjà rencontré dans votre enfance : trois missionnaires et trois cannibales doivent traverser une rivière, mais leur barque ne peut contenir que deux personnes. Si à un moment les cannibales (affamés) sont plus nombreux d'un côté de la rivière que les (appétissants) missionnaires… bon, vous voyez probablement ce qui se passe. Comment tous les six arrivent-ils à traverser la rivière sains et saufs ?

Les créateurs du General Problem Solver, les chercheurs en sciences cognitives Herbert Simon et Allen Newell, avaient enregistré plusieurs étudiants qui « réfléchissaient à voix haute » pendant qu'ils résolvaient ce problème et d'autres énigmes logiques. Simon et Newell conçurent alors leur programme de manière à imiter ce qu'ils estimaient être les processus de pensée des étudiants.

Je ne vais pas entrer dans les détails du fonctionnement du GPS, mais on peut percevoir sa nature symbolique en regardant la manière dont

les instructions de ce programme étaient codées. Dans le langage du GPS, le codage du problème par un humain ressemblerait à peu près à ceci :

ÉTAT ACTUEL :
RIVE-GAUCHE = [3 MISSIONNAIRES, 3 CANNIBALES, 1 BARQUE]
RIVE-DROITE = [VIDE]

ÉTAT SOUHAITÉ
RIVE-GAUCHE = [VIDE]
RIVE-DROITE = [3 MISSIONNAIRES, 3 CANNIBALES, 1 BARQUE]

En français, ces lignes représentent le fait qu'initialement, la rive gauche de la rivière « contient » trois missionnaires, trois cannibales et une barque, tandis que la rive droite ne contient rien. L'état souhaité représente le but du programme – amener tout le monde sur la rive droite de la rivière.

À chaque étape de cette procédure, le GPS tente de modifier l'état courant pour le rapprocher de l'état souhaité. Dans son code, le programme contient des « opérateurs » (sous forme de sous-programmes) qui peuvent transformer l'état courant en un nouvel état et des « règles » qui codent les contraintes associées à la tâche. Il existe ainsi un opérateur qui déplace un certain nombre de missionnaires et de cannibales d'un bord à l'autre de la rivière :

DÉPLACE (#MISSIONNAIRES, #CANNIBALES, DU-BORD, AU-BORD)

Les mots entre parenthèses s'appellent des arguments, et lorsque le programme tourne, il remplace ces mots par des nombres ou d'autres mots. Autrement dit, #MISSIONNAIRES est remplacé par le nombre de missionnaires à déplacer, #CANNIBALES est remplacé par le nombre de cannibales à déplacer, et DU-BORD et AU-BORD sont remplacés par « RIVE-GAUCHE » ou « RIVE-DROITE » en fonction de la rive dont les missionnaires et les cannibales doivent être déplacés. Le code du programme « sait » que la barque se déplace avec les missionnaires et les cannibales.

Avant de pouvoir appliquer l'opérateur « DEPLACE » sur des valeurs spécifiques qui remplacent les arguments, le programme doit vérifier les règles qu'il contient. Par exemple, l'opérateur se bloque si jamais le nombre de personnes dans la barque est supérieur à deux, ou s'il s'avère que son utilisation entraînera sur une rive un surnombre de cannibales par rapport aux missionnaires.

Si ces symboles représentent des concepts – *missionnaires*, *cannibales*, *barque*, *rive gauche* – interprétables par un être humain, l'ordinateur qui exécute ce programme ignore bien sûr tout du sens de ces symboles. Vous pouvez remplacer toutes les occurrences de « MISSIONNAIRES » par « Z372B » ou toute autre suite quelconque de caractères, le programme fonctionnera exactement de la même manière. C'est en partie ce à quoi *General* fait référence dans General Problem Solver. Pour l'ordinateur, le « sens » des symboles résulte de la manière dont on peut les combiner, les lier entre eux, les utiliser.

Les partisans de l'approche symbolique de l'IA soutenaient qu'il n'était pas nécessaire d'écrire des programmes imitant le cerveau pour parvenir à l'ordinateur intelligent et qu'un programme de traitement des symboles suffirait. Certes, le fonctionnement d'un tel programme serait bien plus complexe que l'exemple des Missionnaires et des Cannibales, mais il reposerait toujours sur des symboles, des combinaisons de symboles et des règles et opérations portant sur des symboles. L'IA symbolique de type GPS a fini par prévaloir durant ses trois premières décennies d'existence, surtout sous la forme de *systèmes experts*, où des experts humains concevaient des règles d'une programmation informatique spécialisée dans le but de résoudre des problèmes dans des domaines restreints, tels que les diagnostics médicaux et la prise de décisions juridiques. Il existe plusieurs branches actives de l'IA qui recourent aujourd'hui encore à l'IA symbolique ; j'en donnerai quelques exemples plus loin, en particulier lorsque j'évoquerai les approches du raisonnement et du sens commun *via* l'IA.

L'IA sub-symbolique : les perceptrons

Si l'IA symbolique fut initialement inspirée par la logique mathématique et la manière dont les gens décrivaient leur processus