SUPERINTELLIGENCE

Nick Bostrom

SUPERINTELLIGENCE

L'édition originale de cet ouvrage a été publiée en 2014 en Grande-Bretagne par Oxford University Press sous le titre *Superintelligence, Paths, Dangers, Strategies*. Cette traduction est publiée avec l'accord d'Oxford University Press. Dunod Éditeur est seul responsable de cette traduction de l'œuvre originale et Oxford University Press n'est pas responsable des erreurs, omissions, inexactitudes ou ambiguïtés de cette traduction, ni de toute autre erreur ou omission, imprécision ou ambiguïté dans cette traduction ou pour toute perte causée par la confiance accordée à cette traduction.

Superintelligence was originally published in English in 2014. This translation is published by arrangement with Oxford University Press. Dunod Éditeur is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

Traduction : Françoise Parot

Conception de la couverture et de la maquette intérieure : Grégory Bricout

Illustration de couverture : Claire Scully

Ouvrage publié avec le concours du CNL CENTRE NATIONAL DU LIVRE

11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-086960-2

La fable inachevée des moineaux…

Il était une fois, à la saison où les oiseaux font leur nid, des moineaux qui se reposaient tranquillement, en gazouillant au crépuscule, après de longs, très longs jours de travail.
– Nous sommes si petits et si faibles, comme la vie nous serait facile si nous avions une chouette pour nous aider à construire tous ces nids.
– C'est sûr, lui répondit son voisin, elle nous aiderait aussi à prendre soin de nos parents et de nos enfants,
– Elle nous donnerait des conseils, et elle surveillerait le chat du coin ajouta le suivant.
Alors Pastus, le doyen de la troupe, dit ceci : « Envoyons des éclaireurs dans toutes les directions pour tenter de trouver une jeune chouette ou même un œuf. Un petit corbeau ferait aussi l'affaire, ou même une petite belette. Ce serait sans doute la meilleure chose qui nous soit jamais arrivée, au moins depuis l'ouverture de la Boutique des Graines à Volonté là-bas derrière ».
Ils se mirent tous à rire et, partout, des moineaux commencèrent à gazouiller à plein poumons.
Seul Scronkfinkle, un moineau borgne et râleur, n'était pas du tout convaincu par ce projet. « Ce sera sûrement notre perte, dit-il… nous devrons réfléchir à la manière de domestiquer les chouettes et de les dresser, avant d'introduire chez nous une telle créature… »
Pastus répliqua alors : « Dresser une chouette… voilà qui semble bien délicat. Ce sera déjà assez difficile de trouver un

œuf. Commençons par là et quand nous serons parvenus à avoir un bébé chouette, nous pourrons réfléchir à la manière de le dresser.

« Il y a quelque chose qui ne va pas dans ce projet », s'exclama Scronkfinkle ; mais ses protestations restèrent sans écho, la troupe s'était déjà mise à l'œuvre pour faire ce qu'avait proposé Pastus.

Seuls deux ou trois moineaux restèrent là. Ils commencèrent à réfléchir à ce qu'il faudrait faire pour dresser et domestiquer une chouette. Ils se rendirent vite compte que Pastus avait raison : c'était un défi trop grand, surtout qu'aucune chouette n'était là pour leur dire comment faire. Pourtant, ils y réfléchir comme ils purent, craignant à tout moment que la troupe revienne avec un œuf de chouette avant qu'ils aient trouvé une solution à leur problème.

On ne sait pas comment ça a fini. Mais l'auteur dédie ce livre à Scronkfinkle et à ceux qui l'ont écouté.

Avant-propos

Dans votre crâne, la chose avec laquelle vous êtes en train de lire. Cette chose, le cerveau humain, a des capacités que les autres espèces n'ont pas. Et ce sont ces capacités-là qui nous permettent d'asseoir notre domination sur la planète. Les autres animaux ont une musculature plus puissante, des griffes plus acérées, mais nos cerveaux sont plus intelligents. Ce petit avantage en intelligence générale nous a permis de développer le langage, la technologie ainsi qu'une organisation sociale complexe. Il s'est accru avec le temps car chaque génération s'est appuyée sur les réussites de celles qui l'ont précédée.

S'il nous arrive un jour de construire une machine dotée d'une intelligence générale qui surpassera celle de l'être humain, cette superintelligence pourrait bien alors devenir très puissante. Et, de la même manière que le sort des gorilles dépend aujourd'hui plus des êtres humains que d'eux-mêmes, le sort réservé à notre espèce dépendra des activités-mêmes de cette machine.

Nous avons, c'est vrai, un avantage : c'est nous qui construisons le truc. En principe, on devrait pouvoir mettre au point une superintelligence qui protègerait les valeurs humaines. Et nous aurions bien entendu de très bonnes raisons de le faire. Mais en pratique, ce « problème du contrôle » (contrôle de ce que cette superintelligence ferait) se révèle bien délicat. Tout se passe comme si nous n'avions qu'une seule chance : une fois construite une machine hostile, elle nous empêcherait de la remplacer ou de modifier ses préférences. Notre destin serait scellé.

Dans ce livre, j'essaie de comprendre les menaces éventuelles que représente une telle machine superintelligente et de voir

comment on pourrait y répondre. Il se peut qu'il s'agisse là du défi le plus important et le plus redoutable auquel l'humanité se soit jamais trouvée confrontée. Et, que l'on parvienne ou non à résoudre cette question du contrôle, c'est probablement le dernier défi que nous devrons relever.

Dans ce livre, je ne me centre pas sur l'idée que nous sommes à la veille d'une rupture capitale dans le domaine de l'intelligence artificielle (IA) ou qu'on peut dire avec précision quand elle se produira. Il est assez probable qu'elle surviendra avant la fin du siècle, mais nous ne le savons pas avec certitude. Les deux premiers chapitres discutent des scénarios possibles et avancent quelques idées sur le déroulement du processus. En fait, ce livre concerne d'abord ce qui se produira après cette rupture. Nous verrons la dynamique d'une explosion de l'intelligence, ses formes et ses pouvoirs ; les choix stratégiques qu'elle pourra faire pour obtenir un avantage décisif. Nous en viendrons ensuite à la question du contrôle et nous nous demanderons comment nous pouvons concevoir les conditions initiales du processus de manière à parvenir à une situation vivable et bénéfique. Vers la fin du livre, nous prendrons de la distance et observerons le tableau général qui aura émergé de notre enquête. Certaines propositions seront avancées sur ce qu'il faudrait faire maintenant pour augmenter nos chances d'éviter plus tard une catastrophe généralisée.

Ce livre n'a pas été facile à écrire. J'espère que la voie que j'ai défrichée permettra à d'autres d'atteindre cette nouvelle ligne de front plus rapidement et plus facilement, qu'ils arriveront alors frais et dispos pour parvenir à étendre encore notre compréhension (et si le chemin que j'ai tracé est quelque peu chaotique et sinueux, j'espère que mes critiques, en jugeant le résultat, ne sous-estimeront pas les dangers que présentait ce terrain avant que je le parcours !).

Ce livre n'a pas été facile à écrire. J'ai essayé d'en faire un livre facile à lire, mais je ne pense pas y être vraiment parvenu. En l'écrivant, je visais des lecteurs un peu moins avancés que moi, et j'ai essayé de réaliser un livre que j'aurais bien aimé lire à leur place. Il se peut qu'il s'adresse en fait à un segment étroit de la population… Pourtant, je pense que le contenu du livre devrait être accessible à beaucoup de lecteurs s'ils acceptent de réfléchir en le lisant et s'ils résistent à la tentation de mal comprendre spontanément chacune des idées nouvelles à cause des clichés dont ils sont nourris. Les non-spécialistes ne doivent pas se décourager devant des précisions mathématiques ou devant le vocabulaire spécialisé : on peut toujours trier pour retenir le point principal et négliger les explications qui l'entourent (inversement, pour les lecteurs qui veulent plus de détails, on peut en trouver beaucoup dans les notes de fin).

Bien des choses que j'ai écrites là sont probablement fausses. Il se peut aussi que je n'ai pas pris en compte certains points, d'une importance capitale, et que cela invalide plus ou moins mes conclusions. J'ai tenu à bien signaler les nuances et les degrés d'incertitude tout au long du texte en répétant à l'envi « éventuellement », « pourrait », « peut-être », « serait capable », « il semble », « très probablement », « presque certain ». Ces termes doivent être pris au sérieux, ils ont été choisis sciemment. Pourtant, ces expressions d'une modestie épistémique ne suffisent pas : il faut leur ajouter la reconnaissance systématique de mon incertitude et de mes défaillances. Il ne s'agit pas de fausse modestie : tout en pensant que mon livre est susceptible d'être réellement faux et inutile, je pense qu'un autre point de vue, qui a été énoncé ici ou là, est totalement ou presque erroné : celui d'une opinion par défaut, ou de « l'hypothèse nulle » selon laquelle on peut pour l'instant ignorer tranquillement et raisonnablement la perspective d'une superintelligence.

Remerciements

La fine pellicule qui entoure l'écriture d'un livre a été relativement perméable. Bien des concepts et des idées qui ont généré cet ouvrage ont émané des conversations qui les ont évoquées ; bien sûr, beaucoup de conceptions venues de l'extérieur pendant que j'écrivais ont été intégrées au texte. J'ai tenté d'être vigilant quant à mes citations, mais tous les travaux qui m'ont influencé étaient trop nombreux pour être documentés.

Pour les conversations à perte de vue qui ont clarifié ma pensée, ma reconnaissance va à beaucoup de monde, parmi lesquels Sam Altman, Dario Amodei, Ross Andersen, Stuart Armstrong, Owen Cotton-Barratt, Nick Beckstead, Yoshua Bengio, David Chalmers, Paul Christiano, Milan Ćirković, Andrew Critch, Daniel Dennett, David Deutsch, Daniel Dewey, Thomas Dietterich, Eric Drexler, David Duvenaud, Peter Eckersley, Amnon Eden, Oren Etzioni, Owain Evans, Benja Fallenstein, Alex Flint, Carl Frey, Zoubin Ghahramani, Ian Goldin, Katja Grace, Roger Grosse, Tom Gunter, J. Storrs Hall, Robin Hanson, Demis Hassabis, Geoffrey Hinton, James Hughes, Marcus Hutter, Garry Kasparov, Marcin Kulczycki, Patrick La Victoire, Shane Legg, Moshe Looks, Willam MacAskill, Eric Mandelbaum, Gary Marcus, James Martin, Lillian Martin, Roko Mijic, Vincent Mueller, Elon Musk, Seán Ó Héigeartaigh, Christopher Olah, Toby Ord, Laurent Orseau, Michael Osborne, Larry Page, Dennis Pamlin, Derek Parfit, David Pearce, Huw Price, Guy Ravine, Martin Rees, Bill Roscoe, Francesca Rossi, Stuart Russell, Anna Salamon, Lou Salkind, Anders Sandberg, Julian Savulescu, Jürgen Schmidhuber, Bart Selman, Nicholas Shackel, Murray Shanahan, Noel Sharkey, Carl Shulman, Peter Singer, Nate Soares, Dan Stoiescu, Mustafa Suleyman, Jaan Tallinn, Alexander Tamas, Jessica Taylor, Max Tegmark, Roman Yampolskiy et Eliezer Yudkowsky.

Pour les commentaires plus précis, j'ai une dette envers Milan Ćirković, Daniel Dewey, Owains Evans, Nick Hay, Keith Mansfield, Luke Muehlhauser, Toby Ord, Jess Riedel, Anders Sandberg, Murray Shanahan et Carl Shulman.
Pour la préparation du manuscrit, je remercie Caleb Bell, Malo Bourgon, Robin Brandt, Lance Bush, Cathy Douglass, Alexandre Erler, John King, Kristian Rönn, Susan Rogers, Kyle Scott, Andrew Snyder-Beattie, Cecilia Tilli et Alex Vermeer. Je suis reconnaissant envers mon éditrice Keite Mansfield pour ses encouragements permanents tout au long du projet. Je prie tous ceux dont je ne me suis pas souvenu ici de m'excuser.
Enfin, je remercie affectueusement ceux qui ont financé ce travail, ainsi que mes amis et ma famille : sans votre soutien, ce livre n'aurait pas existé.

1

Ce qui est déjà acquis et ce que nous saurons faire

Commençons par le passé : l'Histoire générale révèle une succession de modes de croissance différents, chacun plus rapide que ceux qui l'ont précédé. Sur la base de ce constat, on peut prévoir un nouveau mode de croissance, donc encore plus rapide. Pourtant, nous n'accorderons pas une grande place à cette conjecture : ce livre ne porte pas sur « l'accélération technologique » ni sur « la croissance exponentielle » ni même sur les diverses conceptions de ce qu'on résume ici ou là par « singularité ». Nous allons donc revenir sur l'histoire de l'intelligence artificielle (IA) puis nous nous interrogerons sur nos capacités actuelles en la matière. Pour finir, nous nous attarderons sur de récentes enquêtes menées auprès d'experts et ferons face à notre ignorance sur le déroulement temporel des progrès à venir.

La croissance dans l'Histoire

Il y a quelques millions d'années seulement, nos ancêtres se balançaient encore dans les branches de la canopée africaine. À l'échelle géologique, ou même évolutive, l'apparition d'*Homo sapiens* à partir de l'ancêtre que nous avons en commun avec les grands singes a été très rapide. On a développé la station debout, le pouce opposable et, de manière décisive, des changements mineurs dans la taille de notre cerveau et de son organisation ont déclenché un bond

capital de nos capacités cognitives : les humains peuvent penser de manière abstraite, communiquer des idées complexes et, bien plus que tout autre espèce de la planète, transmettre des connaissances de génération en génération grâce à la culture.

Ces capacités ont permis aux êtres humains de développer des techniques efficaces de plus en plus nombreuses, et nos ancêtres purent par exemple se déplacer loin de la forêt équatoriale ou de la savane. Après l'invention de l'agriculture en particulier, la densité de population a augmenté en même temps que le nombre total d'humains sur Terre. Plus d'êtres humains, c'est plus d'idées ; une forte densité démographique, c'est une diffusion plus rapide de ces idées et la possibilité, pour certains individus, de se consacrer au développement d'aptitudes spécialisées. L'ensemble de ces facteurs a augmenté *le taux de croissance de la productivité économique et de la capacité technique*. Ce qui s'est passé plus tard, au moment de la Révolution industrielle, a constitué une deuxième étape dans l'évolution de ce taux de croissance.

Ces modifications du taux de croissance ont eu des conséquences très importantes : il y a quelques centaines de milliers d'années, au début de la préhistoire des hominidés, la croissance (technique) était si lente qu'il a fallu près d'un million d'années pour que la productivité économique croisse suffisamment pour nourrir un million d'individus. Environ 5 000 ans av. J.-C., à la suite de la Révolution agraire, ce taux a augmenté au point que le doublement de la population n'a pris que deux siècles. Et aujourd'hui, à la suite de la Révolution industrielle, la croissance économique mondiale est multipliée par deux toutes les 90 minutes[1].

Même si le taux de croissance actuel se maintenait sur la durée, il produirait des résultats impressionnants. Mais si l'économie mondiale continuait de croître au même

rythme que durant le dernier demi-siècle, le monde serait 4,8 fois plus riche en 2050 et 34 fois plus riche en 2100 qu'aujourd'hui[2].

Mais la perspective d'une croissance exponentielle continue n'est rien à côté de ce qui se produira si le monde connaît encore un autre changement radical du taux de croissance, comparable à ceux qu'ont déclenché la Révolution agraire et la Révolution industrielle. L'économiste Robin Hanson estime, sur la base de l'histoire de l'économie et de données démographiques, que l'économie mondiale a doublé en 224 000 ans lors du Pléistocène, quand nous étions chasseurs-cueilleurs, en 909 ans après l'apparition de l'agriculture, et en 6,3 ans dans la société industrielle (dans le modèle de Hanson, notre époque est un mélange de modes de croissance agricole et industriel et l'économie globale ne double pas encore en 6,3 ans)[3]. Si nous passons à un autre mode de croissance, et s'il est en puissance comparable aux deux précédents, notre nouveau régime de croissance verra l'économie mondiale doubler en taille toutes les deux semaines environ.

Aujourd'hui, un tel taux de croissance nous semble fantastique. Dans le passé, il est probable que des observateurs auraient jugé très farfelu de prévoir que l'économie mondiale doublerait un jour plusieurs fois au cours d'une vie. Et pourtant c'est bien dans cette situation extraordinaire que nous nous trouvons aujourd'hui.

La conviction que va se produire une *singularité technologique* est aujourd'hui largement répandue, depuis l'essai fondateur de Vernor Vinge et les travaux qui suivirent, comme ceux de Ray Kurzweil et de quelques autres[4]. Ce terme, « singularité », a néanmoins été utilisé de manière confuse dans bien des usages et a apporté une contribution regrettable à tout un ensemble d'idées techno-utopistes[5]. La plupart de ces sens et de ces idées n'ont aucune importance pour notre propos, aussi gagnerons-nous en clarté en nous

dispensant du terme de « singularité » et en lui préférant une terminologie plus précise.
La seule chose qui soit liée dans ce livre à cette idée d'une singularité technologique est la possibilité d'une explosion de l'intelligence, et précisément la perspective de l'invention d'une machine superintelligente. Il y en a sûrement qui sont convaincus par la courbe de croissance représentée sur la figure 1 et qui pensent qu'un autre changement drastique du mode de croissance est dans les tuyaux, comparable à ceux de la Révolution agraire et de la Révolution industrielle. Ceux-là doivent donc penser qu'il est difficile de concevoir un scénario dans lequel l'économie mondiale pourrait en venir à doubler en quelques semaines sans qu'aient été créés des esprits plus rapides et plus efficaces que ceux de notre espèce biologique. Cependant, pour prendre au sérieux la perspective d'une révolution de l'intelligence des machines, il n'est pas nécessaire de tenir compte des exercices de projection des courbes ou d'extrapolations à partir des croissances économiques antérieures. Comme nous allons le voir, il y a des raisons bien plus fortes.

Les grandes espérances

Depuis qu'ont été inventés les ordinateurs dans les années 1940, on a attendu des machines qu'elles égalent les humains en intelligence *générale*, c'est-à-dire en capacité d'apprendre, de raisonner, de se confronter à tout un ensemble de défis, de traiter des informations complexes dans des domaines matériels comme abstraits. À l'époque, on prévoyait souvent que de telles machines seraient réalisées d'ici une vingtaine d'années[7]. Depuis, la date de cette réalisation a reculé au rythme d'une année tous les ans : et aujourd'hui, les futuristes qui s'intéressent à la possibilité d'une intelligence artificielle générale croient encore souvent que la machine qui en sera capable sera produite d'ici deux ou trois décennies[8].

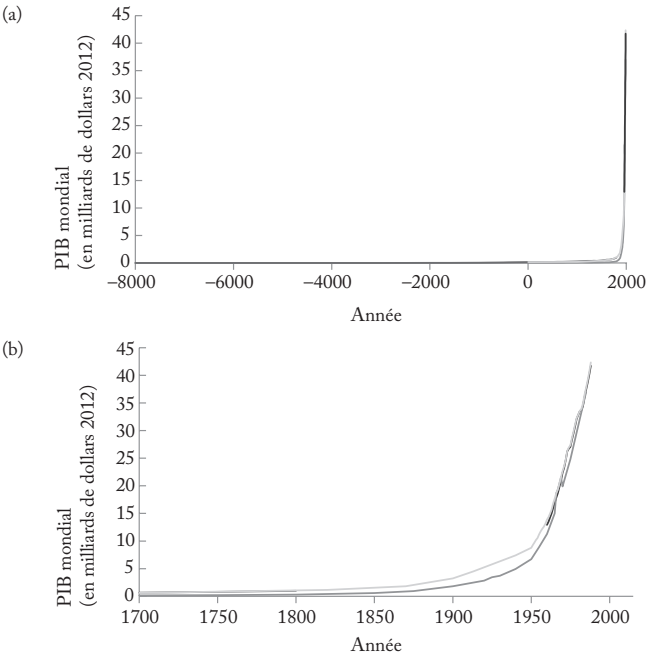

Figure 1 Histoire du produit brut mondial (exprimé en milliards de dollars 2012) sur le long terme. Sur la courbe linéaire, l'histoire de l'économie mondiale ressemble à une courbe plate longeant l'axe des abscisses, jusqu'à un pic vertical. Sur la courbe a, même lorsqu'on se concentre sur les dernières 10 000 années, le graphique montre un angle droit de 90°. Sur la courbe b, ce n'est que dans les 100 dernières années environ que la courbe s'élève de manière nette au-dessus du niveau 0. (Les différences entre les lignes correspondent aux sources de données qui diffèrent légèrement[6].)

Deux ou trois décennies, c'est un délai qui convient bien pour ceux qui font le pronostic d'un changement radical : il est suffisamment proche de nous pour attirer notre attention et notre intérêt, mais suffisamment lointain pour que nous fassions

l'hypothèse d'une série de percées que nous ne parvenons que vaguement à imaginer et qui pourraient se produire d'ici là. On peut comparer ce délai à d'autres, plus courts : la plupart des technologies qui pourraient avoir un impact important sur le monde d'ici cinq ou dix ans sont aujourd'hui disponibles pour un usage limité, et des technologies qui reformateraient le monde d'ici quinze ans maximum existent déjà comme prototypes dans des laboratoires. Et puis vingt ans, c'est une durée qui se rapproche du temps qui reste le plus souvent aux prévisionnistes pour leur carrière, ce qui protège leur réputation d'une prédiction trop risquée.

Mais si certains ont surestimé dans le passé l'apparition d'une telle intelligence artificielle, il ne faut pas en conclure qu'elle est impossible ou qu'elle ne sera jamais mise au point[9]. La principale raison de la lenteur des progrès, c'est que les pionniers en la matière avaient sous-estimé les difficultés techniques de la construction de ce genre de machine. Et cela laisse ouverte la question de l'ampleur de ces difficultés et du temps qu'il faudra pour en venir à bout. Il arrive qu'un problème qui paraissait insurmontable trouve une solution d'une simplicité imprévue (même si l'inverse est sans doute plus fréquent).

Dans le chapitre suivant, nous nous intéresserons aux divers scénarios qui pourraient mener à une machine dont l'intelligence serait égale à celle de l'homme. Mais commençons par remarquer que, malgré les nombreuses étapes qui nous séparent encore de ce genre de machine, elle n'est pas le but final : la station suivante sur la ligne, à courte distance, est une machine *bien plus* intelligente que l'homme. Le train ne marquera pas d'arrêt ou ne ralentira pas à la gare d'Humanville. Il sifflera juste en passant.

Le mathématicien I. J. Good, qui était le statisticien en chef dans l'équipe d'Alan Turing au moment du décryptage pendant la Seconde Guerre mondiale, a peut-être été le premier à énoncer les aspects essentiels de ce scénario :

> « Supposons qu'existe une machine surpassant en intelligence tout ce dont est capable un homme, aussi brillant soit-il. La conception de ce genre de machine faisant partie des activités intellectuelles, cette machine pourrait à son tour créer des machines plus puissantes qu'elle-même ; cela aurait sans nul doute pour effet une "explosion d'intelligence", et l'intelligence humaine resterait loin derrière. La première machine superintelligente sera donc la dernière invention que l'homme aura besoin de faire lui-même, à condition que ladite machine soit assez docile pour nous dire comment la garder sous notre contrôle »[10].

Il devrait être évident aujourd'hui qu'une telle explosion d'intelligence ferait courir des risques majeurs à la vie humaine, et cette perspective devrait être étudiée avec le plus grand sérieux même si l'on pouvait penser (ce qui n'est pas le cas) qu'il n'y a qu'une faible probabilité pour que cela se produise. La plupart des pionniers de l'intelligence artificielle, en dépit de leur conviction selon laquelle nous parviendrons à concevoir une machine de niveau humain très rapidement, n'envisagent pas la possibilité d'une machine très supérieure à l'homme. C'est comme si leur muscle de l'imagination s'était épuisé à concevoir la possibilité radicale de machines atteignant le niveau humain ; ils ne parviennent pas à faire le pas suivant, à savoir que les machines deviendront ensuite superintelligentes.

La plupart d'entre eux ne veulent pas admettre la possibilité que leur travail puisse nous faire courir des risques[11]. Ils ne témoignent d'aucun intérêt (et encore moins d'une réflexion sérieuse) pour tout souci de sécurité et n'éprouvent aucun scrupule éthique quant à la création d'esprits artificiels et à l'éventuelle suprématie d'un ordinateur ; et c'est une lacune qui ne cesse pas d'étonner, même quand on connaît les critères peu exigeants de l'évaluation des technologies[12]. Il faut espérer qu'au moment où la conception d'une telle machine

sera à portée de main, nous aurons acquis les compétences nécessaires pour déclencher une explosion d'intelligence, mais aussi et surtout que nous en aurons la maîtrise suffisante pour survivre à la détonation.
Mais avant de nous tourner vers ce qui nous attend, voyons ce qu'il en a été de l'histoire des artefacts intelligents jusqu'à maintenant.

Un temps pour espérer, un temps pour se décourager

Durant l'été 1956 au Dartmouth College, dix savants qui partageaient leurs intérêts pour les réseaux neuronaux, la théorie des automates et l'étude de l'intelligence se retrouvèrent pour un séminaire de six semaines. On considère souvent que le *Dartmouth Summer Project* est la naissance même de l'intelligence artificielle comme domaine de recherche. Nombre de ceux qui y participèrent furent plus tard reconnus comme ses pères fondateurs. L'optimisme fondamental des délégués se reflète dans la proposition qui fut soumise à la Fondation Rockefeller, qui finança l'événement :

> « Nous nous proposons de soutenir un travail mené par dix hommes pendant deux mois sur l'intelligence artificielle… Il a pour base l'hypothèse que chaque aspect de l'apprentissage ou de quelque autre caractéristique de l'intelligence peut en principe être décrit avec tant de détails qu'une machine pourra être construite pour la simuler. On tentera de découvrir comment fabriquer des machines qui utiliseront le langage, formeront des abstractions et des concepts, résoudront les problèmes aujourd'hui réservés aux êtres humains et sauront s'améliorer elles-mêmes. Nous considérons que des progrès significatifs peuvent être réalisés dans l'un ou l'autre de ces domaines si un groupe de savants soigneusement sélectionnés travaillent ensemble pendant un été. »

Durant les six décennies qui suivirent ces débuts mouvementés, le domaine de l'intelligence artificielle a traversé des périodes de grandes espérances et d'autres, d'échecs et de découragements.

La première des périodes très excitantes, qui avait commencé avec la rencontre de Dartmouth, fut décrite plus tard par John McCarthy (le principal organisateur de cet événement) comme la période du « Regarde, maman, sans les mains ! ». Pendant ces jours-là, les chercheurs construisirent des systèmes pour réfuter les déclarations du type « Aucune machine ne parviendra jamais à faire *ça* ! », déclarations sceptiques très fréquentes à l'époque. Pour les contrer, les chercheurs en intelligence artificielle créèrent des petits systèmes qui faisaient *ça* dans un « micromonde » (un domaine bien précis, limité, qui permet une version minimale de la performance à réaliser), apportant donc une réponse convaincante et montrant que *ça* peut être réalisé en principe par une machine. Un système de cet ordre, le « *Logic Theorist* », permettait de démontrer la plupart des théorèmes du deuxième chapitre du livre *Principia Mathematica* de Whitehead et Russell et parvenait même à une démonstration beaucoup plus élégante que l'originale, réfutant par-là que les machines ne pouvaient « penser que numériquement » et apportant la preuve qu'elles étaient aussi capables de déductions et de démonstrations logiques[13]. Un programme ultérieur, le *General Problem Solver*, était en principe capable de résoudre un grand nombre de problèmes considérés comme formels[14]. On mit au point aussi des programmes qui pouvaient résoudre les problèmes de calcul comme ceux qu'on étudie dans les premières années d'université, les problèmes d'analogie visuelle comme ceux qui figurent dans les tests d'intelligence comme le QI et les problèmes d'algèbre simples[15]. Le robot Shakey (baptisé ainsi à cause de sa tendance à trembloter) montra comment le raisonnement logique pouvait être intégré à la perception et utilisé pour planifier et contrôler

l'activité physique[16]. Le Programme ELIZA montrait comment un ordinateur pouvait simuler un thérapeute rogérien[17]. Au milieu des années 1970, le programme SHRDLU commandait un bras virtuel dans un monde de blocs géométriques en suivant les instructions d'un opérateur et en répondant à ses questions[18]. Au cours des décennies suivantes, on créa des systèmes grâce auxquels des machines pouvaient composer de la musique dans le style de divers compositeurs classiques, faire mieux que de jeunes médecins dans certaines tâches de diagnostic clinique, conduire des voitures de manière autonome et faire des inventions méritant un brevet[19]. Il y eut même un artefact intelligent qui pouvait faire de bonnes blagues[20] (son humour était… moyen : « qu'est-ce que tu obtiens quand tu croises un œil avec un objet mental ? Une *eye-dea* » ; mais les enfants trouvaient ses jeux de mots rigolos).

Les méthodes qui se sont révélées efficaces dans les premières démonstrations ont souvent été difficiles à généraliser à une large variété de problèmes ou à des problèmes plus difficiles. L'une des raisons est due à « l'explosion combinatoire » des possibilités qui doivent être explorées par des méthodes qui reposent sur une sorte de recherche exhaustive. De telles méthodes fonctionnent pour les exemples simples d'un problème, mais échouent quand les choses se compliquent. Par exemple, pour démontrer un théorème en 5 lignes au sein d'un système de déduction qui comprend 1 règle d'inférence et 5 axiomes, on peut se contenter d'énumérer les 3 125 combinaisons possibles et voir si chacune d'elles parvient à la conclusion attendue. Et ça marche aussi avec une démonstration en 6 ou en 7 lignes. Mais au fur et à mesure que la tâche devient plus compliquée, la méthode de la recherche exhaustive rencontre des problèmes. Démontrer un théorème en 50 lignes ne prend pas 10 fois plus longtemps que quand la démonstration en nécessite 5 : en fait, si l'on utilise la recherche exhaustive, cela suppose de combiner $5^{50} \approx 8{,}9 \times 10^{34}$ séquences possibles ; ce

qui est computationnellement impossible même avec les plus rapides des *super-computers*.

Pour éviter cette explosion combinatoire, on a besoin d'algorithmes qui exploitent la structure du domaine cible, qui surpassent la connaissance initiale et recourent à une recherche heuristique, à une planification et à des représentations abstraites, toutes capacités qui n'étaient pas beaucoup développées dans les premiers systèmes d'intelligence artificielle. La performance de ces derniers pâtissait aussi beaucoup des méthodes pauvres de prise en compte de l'incertitude, de la faible fiabilité des représentations symboliques sur lesquelles ils s'appuyaient, de l'insuffisance des données et d'importantes limites de capacité de mémoire et de vitesse de travail des processeurs. Au milieu des années 1970, on commença à faire plus attention à ces problèmes : on réalisa que nombre des projets de l'intelligence artificielle ne pourraient pas tenir leurs promesses initiales et cela a mené au premier hiver de l'intelligence artificielle, c'est-à-dire à une période de recul au cours de laquelle les financements s'amenuisèrent alors que le scepticisme montait ; l'intelligence artificielle cessa d'être à la mode.

Un nouveau printemps survint au début des années 1980, lorsque le Japon lança le projet « Ordinateurs de la 5e génération », une coopération grassement financée par le secteur public et le secteur privé pour dépasser la situation en développant massivement une architecture informatique de travail en parallèle qui pourrait servir de base à l'intelligence artificielle. Ce programme arrivait au moment où l'on parlait du « miracle économique japonais », et les gouvernements occidentaux comme les leaders économiques tentaient de deviner par quelle formule magique ce succès économique avait été déclenché dans l'espoir de répéter cette formule chez eux. Et quand le Japon décida d'investir dans l'intelligence artificielle, les autres pays en firent autant.

Dans les années qui suivirent, on assista donc à une prolifération considérable de *systèmes experts*. Mis au point pour aider à la prise de décision, ces systèmes étaient à base de règles qui permettaient des inférences simples à partir d'une base de connaissances de faits, déterminée d'après des experts humains d'un domaine et minutieusement encodée à la main dans un langage formel. Des centaines de systèmes experts furent élaborés. Mais les petits systèmes apportaient peu de progrès, et les plus grands nécessitaient beaucoup d'argent pour les développer, les valider, les mettre constamment à jour et leur utilisation était en général pénible. Ce n'était pas pratique d'acquérir un ordinateur autonome juste pour faire tourner un seul programme. À la fin des années 1980, le temps de la croissance prit fin.

Ce projet de la 5e génération ne parvint pas à réaliser ses objectifs, et les États-Unis et l'Europe non plus. Survint alors un second hiver. Un critique aurait eu toutes les raisons de se lamenter : « Jusqu'à maintenant, la recherche en intelligence artificielle n'a jamais récolté que des succès limités dans des domaines particuliers suivis immédiatement d'échecs devant des domaines plus étendus auxquels les succès du début laissaient penser qu'on parviendrait. Les investisseurs privés commencèrent à éviter de prendre des risques dans les entreprises impliquées dans l'intelligence artificielle »[21]. Et même les chercheurs académiques et les financements des recherches cessèrent d'employer cette expression[22].

Le travail technique a continué pourtant sans relâche et, dans les années 1990, le dégel mit fin au second hiver de l'intelligence artificielle. L'optimisme reprit des couleurs grâce à l'introduction de nouvelles techniques qui semblaient offrir d'autres possibilités que le paradigme logiciste traditionnel (qu'on appelle souvent la Bonne Vieille Intelligence Artificielle – en anglais « GOFAI ») qui s'était concentré sur la manipulation de symboles de haut-niveau et avait atteint son apogée dans les

systèmes experts des années 1980. Les nouvelles techniques, qui comprenaient les algorithmes génétiques et les réseaux neuronaux, promettaient de surmonter certains des défauts de cette approche GOFAI, en particulier la « fragilité » caractéristique des programmes classiques (qui produisaient des non-sens complets si les programmeurs faisaient ne serait-ce qu'une seule supposition erronée). Les nouvelles techniques se vantaient d'une performance de type plus organique : par exemple, les réseaux neuronaux avaient aussi la propriété de « dégradation contrôlée » en vertu de laquelle une légère atteinte à un réseau neuronal donnait lieu à une dégradation proportionnée de ses performances et non à un crash généralisé. Plus important encore, ces réseaux neuronaux pouvaient apprendre à partir de leur expérience, en trouvant des manières naturelles de faire des généralisations à partir d'exemples et de découvrir les patterns statistiques cachés dans les données[23]. C'est ce qui rendait ces réseaux efficaces dans la reconnaissance de patterns et dans la classification des problèmes : ainsi, en apprenant à un réseau neuronal un ensemble de signaux sonar, il devenait capable de distinguer mieux que des experts humains les profils acoustiques des sous-marins, des mines et d'organismes marins, et cela sans qu'il ait été auparavant nécessaire d'anticiper exactement comment ces profils allaient être élaborés ni comment des traits différents devaient être pondérés.

Certes les réseaux neuronaux simples étaient connus depuis la fin des années 1950, mais ce domaine connut une renaissance après l'introduction des algorithmes de rétropropagation du gradient, qui permettent un apprentissage dans un réseau neuronal multicouches[24]. Ce type de réseau, qui peut comporter une ou plusieurs couches intermédiaires (cachées) entre les couches d'input et d'output, peut apprendre un ensemble beaucoup plus grand de fonctions que leurs prédécesseurs[25]. Associée à des ordinateurs de plus en plus puissants, cette amélioration des algorithmes permit aux

ingénieurs de construire des réseaux neuronaux suffisant pour être utilisés dans beaucoup d'applications.

Les ressemblances entre ces réseaux neuronaux et le cerveau dépassèrent tellement la rigidité tatillonne mais très fragile des systèmes GOFAI traditionnels qu'on forma un nouveau mot en « isme », *connexionnisme*, qui mettait l'accent sur une architecture sub-symbolique massivement parallèle. Plus de 150 000 articles scientifiques ont depuis été publiés sur les réseaux de neurones artificiels et cette approche reste importante dans l'apprentissage automatique.

L'émergence de méthodes évolutives, comme les algorithmes et les programmes génétiques, a également contribué à mettre un terme au second hiver de l'intelligence artificielle. Peut-être cette approche a-t-elle eu un moindre impact dans le milieu académique que les réseaux neuronaux mais elle a été largement médiatisée. Dans les modèles évolutifs, une population de solutions (ce peut être des structures de données ou des programmes) est maintenue, et de nouvelles solutions sont générées aléatoirement en mutant ou recombinant des variantes de cette population initiale. Périodiquement, la population est soumise à un tri sélectif (sur la base de la « fonction fitness », ou fonction d'évaluation de l'adaptation) qui ne retient, dans la génération suivante, que les meilleures. À force de répéter cette procédure des milliers de fois, on accroît étape par étape la qualité moyenne de la population de solutions. Quand cela fonctionne, ce type d'algorithme engendre des solutions efficaces pour un ensemble très large de problèmes ; ces solutions peuvent être radicalement originales et non intuitives, et elles ressemblent souvent plus aux structures naturelles que celles que tout ingénieur pourrait mettre au point. Et en principe, cela n'implique pas plus d'apports que la spécification initiale de la fonction fitness, qui est souvent très simple. Cependant en pratique, recourir à des méthodes évolutives pour avoir de bons résultats nécessite des aptitudes et de l'ingéniosité, en

particulier pour déterminer le bon format représentationnel. Sans une procédure efficace d'encodage des solutions candidates (un langage génétique qui fait correspondre la structure latente au domaine cible), la recherche évolutive s'enfonce sans cesse dans les méandres d'un vaste espace ou reste coincée sur un *optimum* local. Et même si l'on trouve un bon format représentationnel, l'évolution a des exigences computationnelles et elle est mise en échec par l'explosion de ces demandes.

Les réseaux neuronaux et les algorithmes génétiques font partie des méthodes qui ont provoqué beaucoup d'agitation dans les années 1990 parce qu'elles semblaient offrir des alternatives au paradigme des GOFAI, qui stagnaient. Mais il ne s'agit pas ici de chanter leurs louanges ou de les mettre au-dessus des nombreuses autres techniques d'apprentissage automatique. En fait, l'un des développements théoriques majeurs des deux dernières décennies a été de bien comprendre que toutes ces techniques disparates et superficielles n'étaient que des cas particuliers relevant toutes d'un cadre mathématique commun. Ainsi plusieurs types de réseaux de neurones artificiels ne sont rien d'autres que des classifieurs qui effectuent un type particulier de calcul statistique (l'estimation du maximum de vraisemblance)[26]. Cette perspective permet de comparer les réseaux neuronaux à une classe plus étendue d'algorithmes pour les classifieurs d'apprentissage à partir d'exemples – entre autres, les « arbres de décision », les « modèles de régression logistique », les « machines à vecteurs de support », les « classifieurs bayésiens naïfs » des « régressions selon la méthode des *k* plus proches voisins »[27]. Et de la même manière, les algorithmes génétiques peuvent être considérés comme des algorithmes « *hill-climbing* » (rarement exprimés en français « escalade »), c'est-à-dire comme un sous ensemble de très nombreux algorithmes d'optimisation. Chacun de ces algorithmes de détermination des classifieurs ou de recherche d'un espace de solutions a ses avantages et ses inconvénients,

qu'on peut étudier mathématiquement. Les algorithmes diffèrent par leur temps d'exécution et leur espace-mémoire, par les biais inductifs qui leur sont propres, par la facilité d'incorporation de contenus externes et par la transparence de leurs opérations pour l'analyste humain.

Sous l'apparence spectaculaire de l'apprentissage automatique et de la résolution créative de problèmes se cache tout un ensemble de compromis mathématiques bien spécifiés ; l'exemple même en est la thèse d'un agent bayésien parfait qui ferait un usage optimal de l'information disponible. C'est un idéal qui ne peut être atteint parce que, pour l'implémenter dans un processeur physique, il faudrait beaucoup trop de calculs (voir encart 1). Or, on peut considérer que ce que cherche l'intelligence artificielle, ce sont des raccourcis : des moyens de s'approcher de l'idéal de l'agent bayésien tout en sacrifiant un peu d'optimalité ou de généralité mais en ne perdant pas la qualité de la performance dans les domaines concernés.

Encart 1 : Un agent bayésien optimal

Un agent bayésien idéal commence avec une « distribution de probabilités à priori », une fonction qui assigne une probabilité à chacun des « mondes possibles » (c'est-à-dire à chaque manière très spécifique qu'aurait le monde d'apparaître)[28]. Cette probabilité tient compte du biais inductif pour que les mondes possibles plus simples aient une probabilité supérieure (l'une des manières de formaliser la simplicité d'un monde est effectuée en termes de « complexité de Kolmogorov », à savoir une mesure fondée sur la longueur du programme informatique le plus court nécessaire pour générer la description complète du monde[29]). La probabilité a priori tient aussi compte des connaissances acquises que les programmeurs veulent conférer à l'agent.

Quand l'agent reçoit de ses capteurs une nouvelle information, il met à jour sa distribution de probabilités en la conditionnant à la nouvelle information selon le théorème de Bayes[30] : il s'agit de l'opération mathématique qui remet à 0 la nouvelle probabilité des mondes qui sont incompatibles avec l'information nouvelle reçue et qui renormalise la distribution de probabilités de ceux qui restent possibles.

Encart 1 (suite)

Le résultat est une « distribution de probabilités a posteriori » (que l'agent peut utiliser comme nouvelle distribution de probabilités a priori à l'étape suivante). Au fur et à mesure que l'agent fait des observations, la densité de probabilité se concentre sur un ensemble réduit de mondes possibles qui restent compatibles avec les données ; et parmi ces mondes possibles, les plus simples sont les plus probables.

On pourrait considérer la probabilité comme du sable sur une grande feuille de papier. Celle-ci est divisée en zones de taille variable, chacune correspondant à un monde possible, avec les zones les plus larges pour les mondes les plus simples. Imaginons une couche de sable d'épaisseur égale sur toute la feuille : c'est la distribution de probabilité a priori. Chaque fois qu'une information arrive qui disqualifie certains mondes possibles, on enlève le sable qui était sur leurs zones et on le répartit sur les zones restantes. La quantité initiale de sable reste toujours la même, elle ne fait que se concentrer progressivement sur certaines zones au fur et à mesure que parviennent des informations. C'est une image de ce qu'est l'apprentissage dans sa forme la plus simple (pour calculer la probabilité d'une *hypothèse*, on mesure la quantité de sable sur chaque zone qui correspond à l'un des mondes possibles dans lesquels cette hypothèse est vraie).

Jusque-là, nous avons défini une règle d'apprentissage. Pour avoir un agent, nous avons en plus besoin d'une règle de décision. Pour cela, nous équipons l'agent d'une « fonction d'utilité » qui assigne un nombre à chacun des mondes possibles. Ce nombre représente la désirabilité du monde selon les préférences de base de l'agent. Maintenant, à chaque étape, l'agent choisit ses actions en maximisant l'utilité attendue[31] (pour définir cette action avec utilité maximale attendue, l'agent peut faire la liste de toutes les actions possibles. Il peut alors calculer la distribution de probabilité conditionnelle à chaque action : c'est la distribution de probabilité qu'on obtient en conditionnant la distribution de probabilité actuelle à l'observation qui vient juste d'être faite avant. Il peut alors calculer la valeur attendue de l'action en faisant la somme de la valeur de chacun des mondes possibles multipliée par la probabilité conditionnelle de ce monde étant donnée cette action[32]).