Chapitre 1

Introduction

Où l'on tente d'expliquer pourquoi l'intelligence artificielle est un sujet méritant d'être étudié et où l'on essaie de cerner sa vraie nature – cette discussion étant essentielle avant d'aller plus loin.

Nous nous qualifions d'*Homo sapiens* (« humains sages ») en raison de l'importance que nous attribuons à notre **intelligence**. Pendant des millénaires, nous avons essayé de comprendre comment nous *pensons* et *agissons*, autrement dit comment notre cerveau, une simple masse organique, peut percevoir, comprendre, prévoir et manipuler un monde bien plus étendu et complexe que lui-même. Le domaine de l'**intelligence artificielle**, ou IA, ne s'intéresse pas uniquement à la compréhension mais aussi à la *construction* d'entités intelligentes, c'est-à-dire de machines capables de calculer comment agir efficacement et de manière sûre dans une grande diversité de situations nouvelles.

Les analystes classent régulièrement l'IA comme l'un des domaines les plus passionnants et dont la croissance est l'une des plus rapides. Elle génère déjà plus de 1 000 milliards de dollars de recettes par an. Kai-Fu Lee, expert en IA, prédit que son impact sera « plus important que tout autre dans l'histoire de l'humanité ». De plus, les frontières de l'IA constituent un champ intellectuel encore largement ouvert. Alors qu'un étudiant d'une discipline plus ancienne comme la physique peut estimer que les meilleures idées ont déjà été découvertes par Galilée, Newton, Curie, Einstein et autres grands scientifiques, l'IA offre encore assez de perspectives pour que des esprits brillants s'y consacrent à plein temps.

À l'heure actuelle, l'IA est composée d'une grande diversité de disciplines allant des plus générales, comme l'apprentissage, le raisonnement et la perception, aux plus spécifiques, comme jouer aux échecs, démontrer des théorèmes mathématiques, écrire des poèmes, conduire un véhicule ou diagnostiquer des maladies. L'IA relève de toutes les tâches intellectuelles : c'est véritablement un domaine universel.

1.1 Qu'est-ce que l'IA ?

Nous venons d'affirmer tout l'intérêt que présente l'IA, mais nous n'avons encore rien dit de sa *nature*. Historiquement, les chercheurs ont développé plusieurs conceptions différentes de l'IA. Certains définissent l'intelligence en termes de fidélité à la performance *humaine*, tandis que d'autres préfèrent une définition abstraite et formelle de l'intelligence appelée **rationalité**, consistant à « faire les bons choix ». La question même de ce qui est en jeu varie également : certains considèrent l'intelligence comme une propriété des processus de pensée et de raisonnement internes, tandis que d'autres se concentrent sur le comportement intelligent en tant que caractérisation externe[1].

1. Aux yeux du profane, les termes « intelligence artificielle » et « apprentissage automatique » sont parfois difficiles à distinguer. L'apprentissage automatique est un sous-domaine de l'IA qui étudie la capacité à améliorer les performances en fonction de l'expérience. Certains systèmes d'IA utilisent des méthodes d'apprentissage automatique pour acquérir des compétences, mais d'autres non.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

De ces deux dimensions – humain *versus* rationnel[2] et raisonnement *versus* comportement – découlent quatre combinaisons possibles, qui ont chacune leurs partisans et leurs programmes de recherche. Les méthodes utilisées sont nécessairement différentes : la quête d'une intelligence de type humain doit relever en partie d'une science empirique liée à la psychologie, impliquant des observations et des hypothèses sur le comportement humain réel et les processus de pensée ; une approche rationaliste, en revanche, implique une conjugaison de mathématiques et d'ingénierie, et se rattache aux statistiques, à la théorie du contrôle et à l'économie. Les différents groupes se sont à la fois dénigrés et entraidés. Examinons les quatre approches plus en détail.

1.1.1 Agir comme un humain : le test de Turing

Le **test de Turing**, proposé par Alan Turing (1950), a été conçu comme une expérience de pensée permettant de contourner le caractère imprécis de la question philosophique « Une machine peut-elle penser ? » Un ordinateur réussit le test quand un interrogateur humain, après avoir posé des questions par écrit, ne peut pas distinguer si les réponses, également écrites, proviennent d'une personne ou d'un ordinateur. Le chapitre 27 étudie ce test en détail et pose la question de la réelle intelligence d'un ordinateur qui le réussirait. Pour l'instant, contentons-nous de noter que programmer un ordinateur pour réussir véritablement ce test demande d'avoir résolu un grand nombre de problèmes. Un tel ordinateur aurait besoin de maîtriser :

- le **traitement du langage naturel**, pour pouvoir communiquer couramment dans une langue humaine ;
- la **représentation des connaissances**, pour mémoriser ce qu'il sait ou entend ;
- le **raisonnement automatisé**, pour répondre aux questions et tirer de nouvelles conclusions ;
- l'**apprentissage**, pour s'adapter à de nouveaux contextes et détecter et extrapoler des schémas.

Turing considérait qu'il n'était pas nécessaire de présenter la simulation *physique* d'une personne pour démontrer l'intelligence. Cependant, d'autres chercheurs ont proposé un **test de Turing complet**, qui exige une interaction avec des objets et des personnes du monde réel. Pour réussir le test de Turing complet, un robot doit alors être en plus doté :

- d'un dispositif de **vision par ordinateur** et de reconnaissance de la parole pour percevoir le monde ;
- d'une capacité **robotique** pour manipuler des objets et se déplacer.

Les six domaines que nous venons de citer constituent la majeure partie de l'IA. Néanmoins, dans la pratique, les chercheurs en IA ont consacré peu d'efforts à réussir le test de Turing, jugeant plus important d'étudier les principes sous-tendant l'intelligence. La quête du « vol artificiel » a réussi lorsque des ingénieurs et des inventeurs ont cessé d'imiter les oiseaux pour utiliser des souffleries et s'intéresser à l'aérodynamique. L'ingénierie aéronautique ne se donne pas pour objectif de mettre au point « des machines qui imitent si parfaitement le vol des pigeons que les pigeons eux-mêmes pourraient s'y tromper ».

1.1.2 Penser comme un humain : l'approche cognitive

Pour pouvoir dire qu'un programme donné pense comme un humain, il faut déjà savoir comment pense un être humain. Il existe trois moyens d'y parvenir :

- l'**introspection** : la tentative de saisir ses propres pensées à mesure qu'elles passent ;
- les **expériences psychologiques** : l'observation du comportement d'autres personnes ;
- l'**imagerie cérébrale** : l'observation du fonctionnement du cerveau.

Dès lors qu'on dispose d'une théorie de l'esprit suffisamment précise, il devient envisageable de l'exprimer sous la forme d'un programme informatique. Si les comportements du programme en termes d'entrées-sorties correspondent à ceux des humains, c'est le signe que certains de ces mécanismes sont également susceptibles d'opérer chez les humains.

Par exemple, Allen Newell et Herbert Simon, qui ont développé le programme GPS (*General Problem Solver*) [Newell et Simon, 1961], ne se sont pas contentés de créer un programme qui résolvait correctement des

2. Nous ne prétendons pas que les humains sont « irrationnels » au sens littéral de « privés d'un sens clair de la raison ». Nous nous contentons de reconnaître que les décisions humaines ne sont pas toujours mathématiquement parfaites.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

problèmes : ils ont aussi cherché à comparer les étapes chronologiques de son raisonnement à celles de sujets humains confrontés aux mêmes problèmes. Le domaine interdisciplinaire des **sciences cognitives** combine les modèles informatiques de l'IA et les techniques expérimentales de la psychologie dans le but d'élaborer des théories précises et vérifiables du fonctionnement de l'esprit humain.

Les sciences cognitives constituent un domaine fascinant en soi, qui justifie pleinement qu'on lui ait consacré plusieurs ouvrages et au moins une encyclopédie (Wilson et Keil, 1999). Nous n'essaierons pas ici de décrire l'état du savoir quant aux processus cognitifs de l'homme. Nous signalerons de temps à autre les similitudes ou les différences entre les techniques de l'IA et la cognition humaine. Ces recherches ne peuvent être qualifiées de scientifiques que si elles font appel à des expérimentations sur des humains ou des animaux. Nous laisserons cela à d'autres ouvrages, car nous supposons ici que le lecteur ne dispose que d'un ordinateur pour mener des expériences.

Aux débuts de l'IA, les deux approches étaient souvent confondues. Les chercheurs avançaient que, si un algorithme accomplissait bien une tâche, alors on pouvait en *conclure* qu'il constituait un bon modèle du fonctionnement de l'esprit humain, ou *vice versa*. Les auteurs contemporains distinguent ces deux types d'assertions ; cette séparation a permis tant à l'IA qu'aux sciences cognitives de se développer plus rapidement. Ces deux disciplines s'enrichissent mutuellement, notamment dans le domaine de la vision, qui incorpore des enseignements de la neuropsychologie aux modèles informatiques. Récemment, la combinaison de méthodes de neuro-imagerie et de techniques d'apprentissage automatique (*machine learning*) pour l'analyse de ces données a conduit à l'apparition d'une capacité à « lire les pensées », c'est-à-dire à déterminer le contenu sémantique des pensées intérieures d'une personne. Cette capacité pourrait, à son tour, apporter un éclairage supplémentaire sur le fonctionnement de la cognition humaine.

1.1.3 Penser rationnellement : les « lois de la pensée »

Le philosophe grec Aristote est l'un des premiers à avoir essayé de codifier le « penser-juste », autrement dit les procédés permettant de raisonner de manière irréfutable. Ses **syllogismes** proposaient des modèles de structures argumentatives qui aboutissaient toujours à des conclusions vraies, dès lors qu'on leur fournissait des prémisses vraies. L'exemple canonique débute par « Socrate est un homme » et « tous les hommes sont mortels » pour conclure que « Socrate est mortel ». Ces lois de la pensée étaient supposées régir les opérations de l'esprit ; leur étude a ouvert le domaine de la **logique**.

Les logiciens du XIX^e^ siècle ont mis au point une notation précise pour les énoncés relatifs à l'ensemble des objets constituant le monde et aux relations qui les lient. (On peut comparer celle-ci avec la notation arithmétique usuelle qui est seulement destinée aux énoncés sur les *nombres*.) Dès 1965, il existait des programmes qui pouvaient, en principe, résoudre *tout* problème soluble formulé en notation logique. En IA, cette tradition, dite **logiciste,** mise sur des programmes de ce genre pour créer des systèmes intelligents.

La logique, au sens conventionnel du terme, exige une connaissance du monde qui soit certaine – condition qui est rarement atteinte dans la réalité. Nous ne connaissons tout simplement pas les règles de la politique ou de la guerre, par exemple, de la même façon que celles des échecs ou de l'arithmétique. La théorie des **probabilités** comble cette lacune, en autorisant un raisonnement rigoureux sur la base d'informations incertaines. En principe, elle permet la construction d'un modèle complet de pensée rationnelle, menant des informations perceptuelles brutes à la compréhension des mécanismes du monde puis à des prédictions sur le futur. Ce qu'elle ne fait pas, c'est générer un *comportement* intelligent. Pour cela, nous avons besoin d'une théorie de l'action rationnelle. La pensée rationnelle, en soi, n'est pas suffisante.

1.1.4 Agir rationnellement : l'approche de l'agent rationnel

Un **agent** est simplement une entité qui agit (« agent » vient du latin *agere*, « faire »). Bien sûr, tous les programmes informatiques calculent quelque chose, mais les agents informatiques sont supposés faire plus : fonctionner de manière autonome, percevoir leur environnement, subsister pendant une période prolongée, s'adapter au changement et créer et poursuivre des objectifs. Un **agent rationnel** est un agent qui agit de manière à atteindre le meilleur résultat ou, dans un environnement incertain, le meilleur résultat espéré.

Dans le cadre d'une approche de l'IA subordonnée aux « lois de la pensée », l'accent est mis sur la validité des inférences. La capacité à élaborer des inférences correctes fait parfois *partie* de la nature d'un agent rationnel,

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

car une façon d'agir rationnellement consiste à déduire logiquement qu'une action donnée est meilleure que les autres, puis à agir conformément à cette conclusion. À l'inverse, il existe des façons d'agir rationnellement en dehors de l'inférence. Par exemple, retirer sa main d'un poêle brûlant est une action réflexe qui est généralement plus efficace qu'une action plus lente décidée après mûre réflexion.

Toutes les qualités requises par le test de Turing permettent également à un agent d'agir rationnellement. La représentation des connaissances et le raisonnement donnent à un agent la faculté de prendre de bonnes décisions. Pour évoluer dans une société complexe, nous devons pouvoir générer des phrases compréhensibles en langage naturel. Nous avons besoin d'apprendre non seulement à des fins d'érudition, mais aussi parce que cela améliore notre capacité à adopter un comportement adéquat, en particulier dans des contextes nouveaux.

L'approche de l'agent rationnel possède deux avantages sur les autres. Premièrement, elle est plus générale que l'approche par les « lois de la pensée », parce que la validité des inférences n'est que l'un des nombreux mécanismes permettant d'accéder à la rationalité. Deuxièmement, elle convient mieux au développement scientifique. La notion de rationalité est mathématiquement bien définie et complètement générale. On peut facilement travailler à partir de cette spécification pour concevoir des agents qui atteignent cette rationalité de manière prouvable, ce qui est totalement impossible si l'objectif est d'imiter le comportement humain ou les processus de pensée.

Pour ces raisons, l'approche de l'agent rationnel a prévalu pendant la plus grande partie de l'histoire de l'IA. Au cours des premières décennies, les agents rationnels étaient construits sur des bases logiques ; ils formaient des plans précis pour atteindre des objectifs spécifiques. Plus tard, des méthodes basées sur la théorie des probabilités et sur l'apprentissage automatique ont permis de créer des agents capables de prendre des décisions en condition d'incertitude pour atteindre le meilleur résultat espéré. Pour résumer, *l'IA s'est concentrée sur l'étude et la construction d'agents qui* ***font les bons choix****.* Le bon choix est ici défini par l'objectif que l'on fournit à l'agent. Ce paradigme général est si omniprésent qu'on peut l'appeler le **modèle standard**. Il prévaut non seulement en IA, mais aussi en automatique[3], où un contrôleur minimise une fonction de coût, en recherche opérationnelle[4], où une politique maximise une somme de récompenses, en statistiques, où une règle de décision minimise une fonction de perte, et en économie, où un décideur maximise l'utilité ou une certaine mesure du bien-être social.

Le modèle standard doit être affiné dans le cas des environnements complexes, où la rationalité parfaite – toujours agir de façon optimale – n'est pas atteignable, tout simplement parce que les besoins en calcul sont trop élevés. Les chapitres 5 et 17 traitent explicitement du problème de la **rationalité limitée**, c'est-à-dire du choix de l'action appropriée lorsqu'on ne dispose pas d'assez de temps pour faire tous les calculs souhaitables. Cependant, la rationalité parfaite reste généralement un bon point de départ pour l'analyse théorique.

1.1.5 Machines bénéfiques

Le modèle standard a guidé utilement la recherche en IA depuis ses débuts, mais il n'est probablement pas le bon modèle à long terme. En effet, il nécessite de fournir à la machine un objectif entièrement spécifié.

Les tâches définies formellement, comme le jeu d'échecs et le calcul du chemin le plus court, s'accompagnent d'un objectif bien déterminé, de sorte que le modèle standard est applicable. Cependant, à mesure qu'on se rapproche du monde réel, il devient de plus en plus difficile de spécifier l'objectif de manière complète et correcte. Par exemple, dans le cas de la conception d'un véhicule entièrement automatisé, dit « véhicule autonome »[5], on pourrait penser que l'objectif est d'atteindre la destination en toute sécurité. Pourtant, conduire sur une route, quelle qu'elle soit, comporte un risque inhérent d'être blessé à cause d'autres conducteurs imprudents, de pannes mécaniques, etc. Il y a un compromis à trouver entre progresser vers la destination et encou-

3. NdT. L'automatique est une science de l'ingénieur qui permet d'automatiser des tâches à l'aide de machines fonctionnant sans intervention humaine. On parle alors de système asservi ou régulé.

4. NdT. Cette discipline d'aide à la décision utilise des modèles mathématiques pour résoudre des problèmes complexes en déterminant la solution optimale. Le qualificatif « opérationnelle » s'explique par le fait que la première application avait trait aux opérations militaires. La dénomination est restée, même si la discipline s'est diffusée à d'autres champs (politique, économie, finance, industrie, etc.).

5. NdT. L'autonomie ne se conçoit que pour un agent doué de volonté. L'industrie automobile utilise le terme d'ADAS (*advanced driver-assistance systems*), qui comprend cinq niveaux d'assistance. Ce qu'on appelle communément « véhicule autonome » correspond à un ADAS de niveau 5.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

rir un risque de blessure. Comment ce compromis doit-il être fait ? En outre, dans quelle mesure peut-on permettre à la voiture d'opérer des choix qui gêneraient les autres conducteurs ? De quelle manière le véhicule doit-il gérer son accélération, sa direction et son freinage pour éviter de secouer le passager ? Il est difficile de répondre *a priori* à ce genre de questions. Elles sont particulièrement problématiques dans le domaine général de l'interaction humain-robot, dont la voiture autonome n'est qu'un exemple.

Le problème de faire concorder ses préférences réelles avec l'objectif qu'on assigne à la machine est appelé **problème d'alignement des valeurs** : les valeurs ou objectifs incorporés dans la machine doivent être alignés sur ceux des humains. Si un système d'IA est développé dans un laboratoire ou dans un simulateur, comme cela a été le cas pour la majeure partie de l'histoire du domaine, il y a une solution facile à un objectif mal spécifié : réinitialiser le système, corriger l'objectif et essayer à nouveau. À mesure que le domaine progresse vers des systèmes intelligents de plus en plus performants et déployés dans le monde réel, cette approche n'est plus viable. Déployer un système qui a un objectif incorrect aura des conséquences préjudiciables. En outre, plus le système sera intelligent, plus les conséquences seront graves.

Pour revenir à l'exemple apparemment sans problème des échecs, imaginons ce qui se passerait si une machine était suffisamment intelligente pour raisonner et agir en dehors des limites de l'échiquier. Dans ce cas, elle pourrait tenter d'augmenter ses chances de gagner en rusant : en hypnotisant ou en faisant chanter son adversaire, ou encore en soudoyant le public pour qu'il fasse du bruit pendant son temps de réflexion[6]. Elle pourrait également tenter de détourner de la puissance de calcul supplémentaire pour elle-même. *Ces comportements ne sont ni « non intelligents » ni « délirants » ; ils sont une conséquence logique du fait que gagner est le* seul *objectif défini pour la machine.*

Il est impossible d'anticiper tous les abus qu'une machine poursuivant un objectif fixé pourrait être capable de commettre. Il y a donc de bonnes raisons de penser que le modèle standard est inadéquat. Nous ne voulons pas de machines intelligentes au sens où elles poursuivent *leurs* objectifs ; nous voulons qu'elles poursuivent *nos* objectifs. Si nous ne pouvons pas transférer parfaitement ces objectifs à la machine, alors nous avons besoin d'une nouvelle formulation qui permette à la machine de poursuivre nos objectifs tout en restant nécessairement dans l'*incertitude* quant à ce qu'ils sont. Lorsqu'une machine sait qu'elle ne connaît pas l'objectif exact, elle est incitée à agir avec prudence, à demander la permission, à en savoir plus sur nos préférences par l'observation et à s'en remettre au contrôle humain. En fin de compte, nous voulons des agents à **bénéfice démontré** pour les humains. Nous reviendrons sur ce sujet en section 1.5.

1.2 Fondements de l'intelligence artificielle

Dans cette section, nous présentons un bref historique des disciplines qui ont apporté des idées, des points de vue et des techniques à l'IA. Comme tout historique, celui-ci se focalise sur un petit nombre de personnes, d'événements et d'idées et laisse de côté beaucoup de choses également importantes. Nous organisons cet historique autour d'une série de questions. Nous ne voulons certainement pas donner l'impression que ces problématiques sont les seules abordées par ces disciplines, ni que celles-ci n'ont que l'IA comme finalité ultime.

1.2.1 Philosophie

- Peut-on utiliser des règles formelles pour tirer des conclusions valides ?
- Comment l'esprit émerge-t-il à partir du cerveau physique ?
- D'où la connaissance provient-elle ?
- Comment la connaissance conduit-elle à l'action ?

Aristote (384-322 av. J.-C.) a été le premier à formuler un ensemble précis de lois régissant la partie rationnelle de l'esprit. Il a développé un système informel de syllogismes produisant des raisonnements valides. En principe, ce système autorise quiconque à tirer mécaniquement des conclusions à partir de prémisses initiales.

6. Dans l'un des premiers livres sur les échecs, Ruy Lopez (1561) écrivait : « Placez toujours l'échiquier de manière à ce que votre adversaire ait le soleil dans les yeux. »

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Raymond Lulle (environ 1232-1315) a conçu *ars magna*, un système de raisonnement qu'il présente dans son ouvrage *Le Grand et Dernier Art* (1305). Il a tenté de matérialiser son système en utilisant un véritable dispositif mécanique : un ensemble de roues en papier qui pouvaient tourner et produire différentes permutations.

Vers 1500, Léonard de Vinci (1452-1519) a imaginé sans la construire une machine à calculer dont de récentes reconstitutions ont montré qu'elle aurait pu fonctionner. La première machine à calculer connue a été fabriquée vers 1623 par le scientifique allemand Wilhelm Schickard (1592-1635). Blaise Pascal (1623-1662) a construit la Pascaline en 1642, qu'il décrit ainsi : «[elle] produit des effets qui approchent plus de la pensée que tout ce que font les animaux ». Gottfried Wilhelm Leibniz (1646-1716) a construit une machine destinée à manier des opérations sur les concepts plutôt que sur les nombres, mais son champ était assez limité. Dans son ouvrage *Leviathan* de 1651, Thomas Hobbes (1588-1679) suggère l'idée d'une machine pensante, un « animal artificiel », selon ses propres mots, en argumentant ainsi : « Qu'est-ce que le cœur, sinon un ressort, et les nerfs, sinon autant de cordes, et les articulations, sinon autant de poulies ? » Il a également suggéré que le raisonnement ressemblait au calcul numérique : « Car la "raison" […] n'est rien d'autre que du "calcul", c'est-à-dire ajouter et soustraire. »

C'est une chose d'avancer que l'esprit fonctionne, au moins en partie, selon des règles logiques ou numériques, et de construire des dispositifs physiques qui émulent certaines de ces règles. C'en est une autre que de défendre l'idée que l'esprit lui-même *est* un système physique de ce type. René Descartes (1596-1650) a été le premier à exposer clairement la distinction entre l'esprit et la matière. Il a fait observer qu'une conception purement physique de l'esprit ne semble laisser que peu de place au libre arbitre : si l'esprit est entièrement régi par des lois physiques, alors celui-ci a autant de liberté qu'une pierre qui « décide » de tomber. Descartes était un partisan du **dualisme** : il considérait qu'une partie de l'esprit humain (l'âme) était en dehors de la nature et soustraite aux lois physiques, alors que les animaux étaient dépourvus de cette qualité duale et pouvaient être considérés comme des machines.

Le **matérialisme** s'oppose au dualisme : cette théorie énonce que les opérations du cerveau se conforment aux lois de la physique et *constituent* l'esprit. Le libre arbitre n'est alors plus que l'aspect sous lequel l'entité qui prend la décision perçoit les choix. Les termes **physicalisme** et **naturalisme** sont également utilisés pour décrire ce point de vue qui s'oppose au surnaturel.

La nature physique de l'esprit qui manipule des connaissances étant établie, le problème suivant consiste à définir la source de la connaissance. Le mouvement **empiriste**, qui a commencé avec le *Novum Organum*[7] de Francis Bacon (1561-1626), est caractérisé par une formule de John Locke (1632-1704) : « Il n'y a rien dans l'entendement qui n'ait d'abord été dans les sens. »

Dans son *Traité de la nature humaine* (Hume, 1739), David Hume (1711-1776) proposait ce qu'on appelle désormais le principe d'**induction**, selon lequel les règles générales sont élaborées à partir de la découverte d'associations répétées entre leurs éléments.

S'appuyant sur les travaux de Ludwig Wittgenstein (1889-1951) et de Bertrand Russell (1872-1970), le célèbre cercle de Vienne (Sigmund, 2017), un groupe de philosophes et de mathématiciens qui se réunissait à Vienne dans les années 1920-1930, a développé la doctrine du **positivisme logique**. Dans cette doctrine, toute la connaissance peut être caractérisée par des théories logiques provenant, *in fine*, d'**énoncés d'observation** qui correspondent à des perceptions sensorielles ; le positivisme logique combine donc le rationalisme et l'empirisme.

La **théorie de la confirmation** de Rudolf Carnap et de Carl Hempel (1905-1997) tente d'analyser l'acquisition de la connaissance à partir de l'expérience. Elle quantifie le degré de croyance à assigner à un énoncé logique en se fondant sur les connexions qu'il entretient avec les observations qui le confirment ou l'infirment. L'ouvrage de Carnap *La Structure logique du monde* (1928) décrit peut-être la première théorie de l'esprit vue comme un processus de calcul.

Le dernier élément de la conception philosophique de l'esprit est le lien entre la connaissance et l'action. Cette question est vitale pour l'IA, car l'intelligence requiert autant d'action que de raisonnement. En outre, ce n'est qu'en comprenant comment les actions sont justifiées que l'on peut découvrir comment construire un agent dont les actions sont justifiables (ou rationnelles).

7. Le *Novum Organum* est une nouvelle version de l'*Organon* (ou instrument de la pensée) d'Aristote.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Dans *De Motu Animalium*, Aristote défend l'idée que les actions sont justifiées par un lien logique entre les objectifs et la connaissance du résultat des actions :

> Mais comment se fait-il que la pensée soit parfois accompagnée d'une action et parfois non, parfois d'un mouvement et parfois non ? Il semble qu'il se passe la même chose lorsqu'on raisonne et qu'on produit des inférences à propos d'objets qui ne changent pas. Cependant, dans ce cas, la fin est une proposition spéculative […] tandis que dans l'autre la conclusion produite par les deux prémisses est une action […]. J'ai besoin d'une couverture ; un manteau est une couverture ; j'ai besoin d'un manteau. Ce dont j'ai besoin, je dois le fabriquer ; j'ai besoin d'un manteau, je dois fabriquer un manteau. Et la conclusion « Je dois fabriquer un manteau » est une action.

Dans l'*Éthique à Nicomaque*, Aristote approfondit ce sujet et propose une méthode :

> Nous délibérons non sur les fins, mais sur les moyens. En effet, ni le médecin ne délibère pour savoir s'il doit guérir, ni l'orateur pour savoir s'il doit persuader […]. Mais, ayant posé en principe la fin, ils examinent comment, c'est-à-dire par quels moyens, elle sera réalisée. Et s'il se révèle possible de l'obtenir par plusieurs moyens, ils examinent par lequel elle le sera le plus facilement et le mieux. Si au contraire elle ne peut être accomplie que par un seul moyen, ils examinent *comment* elle sera obtenue par ce moyen, et *ce moyen lui-même*, par quel moyen on l'obtiendra, jusqu'à ce qu'ils arrivent à la première cause, […] et ce qu'on trouve en dernier lieu dans l'ordre de l'analyse, c'est ce qu'on fait en premier lieu dans l'ordre de réalisation[8].

L'algorithme suggéré par Aristote a été implémenté 2 300 ans plus tard par Newell et Simon dans leur programme GPS. De nos jours, on le qualifierait de système de planification par régression (voir chapitre 11). Les méthodes basées sur la planification logique pour atteindre des objectifs bien définis ont dominé les premières décennies de recherche théorique en IA.

Penser uniquement en termes d'actions qui réalisent des objectifs est souvent utile, mais parfois inapplicable. Par exemple, s'il existe plusieurs façons différentes d'atteindre un objectif, il faut pouvoir opérer un choix. Plus important encore, il arrive qu'il ne soit pas possible d'atteindre un objectif avec certitude, mais qu'il faille quand même agir. Comment décider, dans ce cas ? Antoine Arnauld (1662), analysant la notion de décision rationnelle dans les jeux de hasard, a proposé une formule quantitative pour maximiser la valeur monétaire attendue du résultat. Plus tard, Daniel Bernoulli (1738) a introduit la notion plus générale d'**utilité** pour appréhender la valeur interne et subjective d'un résultat. La notion moderne de prise de décision rationnelle sous incertitude suppose la maximisation de l'utilité espérée, comme l'explique le chapitre 16.

En matière d'éthique et de politique publique, un décideur doit prendre en compte les intérêts de nombreuses personnes. Jeremy Bentham (1823) et John Stuart Mill (1863) ont promu l'idée d'**utilitarisme**, selon laquelle la prise de décision rationnelle basée sur la maximisation de l'utilité doit s'appliquer à toutes les sphères de l'activité humaine, y compris aux décisions de politique publique prises au nom du plus grand nombre. L'utilitarisme est un cas particulier de **conséquentialisme**, une théorie qui défend l'idée que ce qui est bien ou mal est déterminé par les effets espérés d'une action.

Emmanuel Kant, quant à lui, propose en 1785 une théorie de l'**éthique déontologique** basée sur des règles, dans laquelle « faire les bons choix » est déterminé non pas par les effets des actions, mais par des lois sociales universelles qui régissent les actions autorisées, comme « ne pas mentir » ou « ne pas tuer ». Ainsi, un utilitariste pourrait proférer un pieux mensonge si le bénéfice attendu l'emporte sur le préjudice, mais un kantien serait tenu de s'en abstenir, car mentir est intrinsèquement mauvais. Mill reconnaissait la valeur des règles, mais les considérait comme de simples procédures de décision efficaces compilées à partir de principes premiers de raisonnement sur les conséquences. De nombreux systèmes d'IA modernes adoptent exactement cette approche.

1.2.2 Mathématiques

- Quelles sont les règles formelles qui permettent de tirer des conclusions valides ?
- Que peut-on calculer ?
- Comment raisonne-t-on à partir d'informations incertaines ?

8. Traduction Gauthier et Jolif, Presses universitaires de Louvain, 1970.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Si les philosophes sont à l'origine de certaines idées fondamentales pour l'IA, la transformation de cette dernière en une véritable science a exigé la mathématisation de la logique et des probabilités, ainsi que l'introduction d'une nouvelle branche des mathématiques : le calcul[9].

On peut faire remonter l'idée de **logique formelle** aux philosophes de l'Antiquité en Grèce, en Inde et en Chine, mais ses développements mathématiques n'ont vraiment commencé qu'avec les travaux de George Boole (1815-1864), qui a défini précisément la logique propositionnelle, ou logique booléenne (Boole, 1847). En 1879, Gottlob Frege (1848-1925) a étendu la logique de Boole afin d'y inclure des objets et des relations. Ce faisant, il a créé la logique du premier ordre que l'on connaît aujourd'hui[10]. En plus de son rôle central aux débuts de la recherche en IA, la logique du premier ordre a motivé les travaux de Gödel et de Turing qui ont fondé le calcul lui-même, comme nous l'expliquons ci-dessous.

La théorie des **probabilités** peut être considérée comme une généralisation de la logique aux situations où les informations sont incertaines – ce qui revêt une grande importance pour l'IA. Jérôme Cardan (1501-1576) est le premier à avoir formulé l'idée de probabilités, en les décrivant à l'aide des issues possibles des événements dans les jeux de hasard. En 1654, Blaise Pascal a montré, dans une lettre à Pierre Fermat (1601-1665), comment prédire la suite d'un jeu de hasard non terminé et comment attribuer des gains moyens aux joueurs. Les probabilités sont rapidement devenues une composante incontournable des sciences quantitatives, permettant de pallier les mesures incertaines et les théories incomplètes. Jacob Bernoulli (1654-1705, oncle de Daniel), Pierre Laplace (1749-1827), et d'autres ont fait progresser la théorie et ont introduit de nouvelles méthodes statistiques. Thomas Bayes (1702-1761) a proposé une règle pour actualiser la théorie des probabilités à la lumière de nouvelles observations ; la règle de Bayes est un outil crucial pour les systèmes d'IA.

La formalisation des probabilités, combinée à la disponibilité des données, a conduit à l'émergence du domaine des **statistiques (émergence)**. L'une des premières utilisations a été l'analyse des données du recensement de Londres de 1662 par John Graunt. Ronald Fisher est considéré comme le premier statisticien moderne (Fisher, 1922). Il a combiné les idées de probabilité, de plan d'expérience, d'analyse des données et de calcul. En 1919, il a fait valoir qu'il ne pouvait pas faire son travail sans une calculatrice mécanique appelée la Millionnaire (la première calculatrice capable de faire des multiplications), même si le coût de la calculatrice était supérieur à son salaire annuel (Ross, 2012).

L'histoire du calcul remonte aussi loin que l'histoire des nombres eux-mêmes, mais le premier **algorithme** non trivial est attribué à Euclide pour sa méthode de calcul des plus grands diviseurs communs. L'origine du mot *algorithme* remonte à Muhammad ibn Musa al-Khawarizmi, mathématicien du IXe siècle dont les écrits ont également introduit les chiffres arabes et l'algèbre en Europe. Boole et d'autres auteurs ont proposé des algorithmes pour la déduction logique et, au cours du XIXe siècle, on s'est efforcé de formaliser des raisonnements mathématiques généraux sous forme de déductions logiques.

Kurt Gödel (1906-1978) a montré qu'il existe une procédure efficace pour prouver tout énoncé vrai de la logique du premier ordre de Frege et de Russell, mais que cette logique ne peut pas rendre compte du principe d'induction mathématique nécessaire à la caractérisation des entiers naturels. En 1931, Gödel a prouvé que la notion de déduction connaît des limites. Avec son **théorème d'incomplétude**, il a montré que, dans toute théorie aussi expressive que celle de l'arithmétique de Peano (la théorie élémentaire des entiers naturels), il existe des énoncés nécessairement vrais qui n'ont pas de preuve dans la théorie elle-même.

Ce résultat fondamental peut aussi s'interpréter comme établissant la preuve qu'il existe des fonctions sur les entiers qui ne peuvent pas être représentées par un algorithme, autrement dit qui ne peuvent pas être calculées. C'est ce qui a incité Alan Turing (1912-1954) à tenter de caractériser avec exactitude les fonctions **calculables** – celles qu'on *peut* calculer par une procédure effective. La thèse de Church-Turing propose d'assimiler la notion générale de calculabilité à l'ensemble des fonctions qui sont calculables par une machine de Turing (Turing, 1936). Turing a aussi montré qu'il existe des fonctions qu'aucune machine de Turing ne peut calculer. Par exemple, aucune machine ne peut dire *en général* si un programme retournera une réponse pour une entrée donnée ou s'il risque de continuer sans jamais s'arrêter.

9. NdT. Le calcul au sens mécanique du terme (*computation*).

10. La notation proposée par Frege pour la logique du premier ordre – une combinaison astucieuse d'éléments textuels et géométriques – n'a jamais connu le succès.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Bien que la calculabilité soit importante pour comprendre la notion de calcul mécanique, celle de **praticabilité** (*tractability*) a eu un impact encore plus grand sur l'IA. On peut la définir approximativement ainsi : un problème est dit impraticable si le temps requis pour en résoudre des exemples croît exponentiellement avec la taille de ces exemples. La distinction entre croissance polynomiale et croissance exponentielle de la complexité a été pour la première fois mise en évidence au milieu des années 1960 (Cobham, 1964 ; Edmonds, 1965). Son importance est liée au fait que même des problèmes de petite taille ne peuvent pas être résolus en un temps raisonnable dès lors que la croissance est exponentielle.

La théorie de la **NP-complétude** élaborée par Cook (1971) et Karp (1972) fournit une méthode pour l'analyse de la praticabilité des problèmes : toute classe de problèmes à laquelle la classe des problèmes NP-complets peut être réduite risque fort d'être impraticable. (Bien qu'on n'ait jamais démontré que les problèmes NP-complets sont nécessairement impraticables, la plupart des théoriciens pensent que c'est le cas.) Ces résultats contredisent l'enthousiasme avec lequel la presse a salué les premiers ordinateurs en les qualifiant de « supercerveaux électroniques » qui étaient « plus rapides qu'Einstein ». Malgré la vitesse croissante des ordinateurs, une utilisation parcimonieuse des ressources et une nécessaire imperfection seront toujours de mise pour les systèmes intelligents. Pour le dire sans détours, le monde est une instance de problème de taille *extrême* !

1.2.3 Économie

- Comment prendre des décisions en accord avec nos préférences ?
- Comment faire quand les autres risquent de ne pas coopérer ?
- Comment y parvenir alors que les gains sont susceptibles d'être éloignés dans le futur ?

Les sciences économiques sont apparues en 1776 avec la publication de la *Recherche sur la nature et les causes de la richesse des nations* d'Adam Smith (1723-1790). Smith y proposait d'analyser les économies comme constituées de nombreux agents individuels veillant à leurs propres intérêts. Toutefois, Smith ne préconisait pas du tout la cupidité financière comme position morale : son livre précédent *La Théorie des sentiments moraux* (1759) commence par souligner que le souci du bien-être d'autrui est une composante essentielle des intérêts de chaque individu.

La plupart des gens pensent que le sujet de l'économie est l'argent. Effectivement, la première analyse mathématique des décisions en situations d'incertitude, la formule de la valeur maximale espérée d'Antoine Arnauld (1662), portait sur la valeur monétaire des paris. Daniel Bernoulli (1738) a remarqué que cette formule ne semblait pas bien fonctionner pour des sommes beaucoup plus importantes, comme les investissements dans les expéditions commerciales maritimes. Il a proposé à la place un principe basé sur la maximisation de l'utilité espérée et il a expliqué les choix d'investissement des humains en avançant que l'utilité marginale d'une quantité additionnelle d'argent diminue au fur et à mesure que l'on gagne plus d'argent.

Léon Walras (1834-1910) a offert un fondement plus général à la théorie de l'utilité en termes de préférences entre paris sur des issues quelconques (pas seulement monétaires). La théorie a été enrichie par Ramsey (1931) et plus tard par John von Neumann et Oskar Morgenstern dans leur ouvrage *Théorie des jeux et comportement économique* (1944). L'économie n'est plus l'étude de l'argent, mais plutôt l'étude des souhaits et des préférences.

La **théorie de la décision**, qui combine la théorie des probabilités et celle de l'utilité, fournit un cadre formel complet pour les décisions individuelles (économiques ou autres) en environnement incertain, autrement dit dans les cas où l'environnement dans lequel on doit prendre la décision peut être représenté par une description probabiliste. Elle convient bien aux économies de « grande dimension » dans lesquelles les agents ne tiennent pas compte des actions des autres agents considérés en tant qu'individus. Pour des économies de petite dimension, la situation ressemble beaucoup plus à celle d'un **jeu** : les actions d'un joueur peuvent exercer une influence considérable sur l'utilité d'un autre (positivement ou négativement). La **théorie des jeux** développée par von Neumann et Morgenstern (voir également Luce et Raiffa, 1957) contient un résultat surprenant : dans certains jeux, un agent rationnel doit adopter une politique aléatoire (ou qui du moins apparaît comme telle). À l'inverse de la théorie de la décision, la théorie des jeux ne procure pas de méthode certaine pour sélectionner les actions. En IA, les décisions mettant en jeu plusieurs agents sont étudiées dans le domaine des **systèmes multiagents** (voir chapitre 18).

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

À de rares exceptions près, les économistes n'ont pas répondu à la troisième question évoquée précédemment, à savoir comment prendre des décisions rationnelles lorsque les perspectives de gains sont éloignées et dépendent de plusieurs actions réalisées *en séquence*. Ce sujet a été étudié dans le domaine de la **recherche opérationnelle**, apparu pendant la Seconde Guerre mondiale à l'occasion des efforts entrepris en Grande-Bretagne pour optimiser les installations radars, avant de trouver d'innombrables applications civiles. Par ses travaux, Richard Bellman (1957) a formalisé une famille de problèmes de décisions séquentielles nommés **processus décisionnels de Markov (*Markov decision processes*)**, qui sont étudiés au chapitre 17 et, sous le titre d'**apprentissage par renforcement (*reinforcement learning*)**, au chapitre 22.

Les travaux en économie et en recherche opérationnelle ont beaucoup contribué à notre notion d'agent rationnel, bien que la recherche en IA se soit longtemps développée en empruntant des chemins complètement distincts. Une des raisons de ce développement propre tient à l'apparente complexité de la prise de décision rationnelle. Herbert Simon (1916-2001), le pionnier de la recherche en IA, a obtenu en 1978 le prix Nobel d'économie pour des travaux montrant que les modèles de prise de décision fondés sur le principe du **seuil de satisfaction (*satisficing*)** de l'individu – les individus sont prêts à accepter de prendre des décisions « suffisamment bonnes » plutôt qu'optimales, évitant ainsi de passer par le calcul laborieux d'un optimum – fournissent une meilleure description du comportement réel des humains (Simon, 1947). Depuis les années 1990, on assiste à une résurgence de l'intérêt à l'égard des techniques de la théorie de la décision en IA.

1.2.4 Neurosciences

- Comment le cerveau traite-t-il l'information ?

Les **neurosciences** étudient le système nerveux et en particulier le cerveau. Bien que la manière exacte dont le cerveau engendre la pensée fasse partie des grands mystères de la science, on admet depuis des millénaires que le cerveau *est*, d'une manière ou d'une autre, à la source de la pensée, pour avoir observé la diminution des capacités mentales consécutive à de violents coups sur la tête. On sait également depuis longtemps que les cerveaux humains sont un peu particuliers ; vers 335 av. J.-C., Aristote écrivait : « De tous les animaux, l'homme a le cerveau le plus important proportionnellement à sa taille »[11]. Ce n'est qu'au milieu du XVIII^e^ siècle que le cerveau fut reconnu comme le siège de la conscience. Auparavant, les emplacements envisagés étaient notamment le cœur et la rate.

En 1861, l'étude par Paul Broca (1824-1880) de l'aphasie (un trouble du langage) chez des patients atteints de lésions cérébrales a initié l'étude de l'organisation fonctionnelle du cerveau par l'identification d'une zone localisée dans l'hémisphère gauche (maintenant appelée aire de Broca) qui est responsable de la production de la parole[12]. On savait déjà que le cerveau est en grande partie composé de cellules nerveuses, ou **neurones**, mais ce n'est qu'en 1873 que Camillo Golgi (1843-1926) a développé une technique de coloration permettant d'observer des neurones individuels (voir figure 1.1). Cette technique a été utilisée par Santiago Ramón y Cajal (1852-1934) dans ses études pionnières sur les structures neuronales[13]. Il est maintenant largement admis que les fonctions cognitives résultent du fonctionnement électrochimique de ces structures. Ainsi, *un ensemble de simples cellules peut engendrer la pensée, l'action et la conscience.* Selon les propres mots de John Searle (1992), *les cerveaux engendrent les consciences*.

On connaît désormais la correspondance entre les zones du cerveau et les parties du corps humain qu'elles contrôlent ou desquelles elles reçoivent des stimuli sensoriels. Ces correspondances peuvent changer complètement en quelques semaines et certains animaux semblent avoir plusieurs systèmes de correspondance. En outre, on ne comprend pas parfaitement comment d'autres zones peuvent compenser les fonctions d'une zone endommagée. Enfin, il n'existe pratiquement aucune théorie sur l'enregistrement de la mémoire individuelle ou sur le mécanisme des fonctions cognitives supérieures.

11. On a découvert depuis que certains petits mammifères (les scandentiens) et certaines espèces d'oiseaux ont un rapport masse du cerveau à masse de l'organisme plus élevé que le nôtre.

12. Alexander Hood (1824) est souvent cité comme un précurseur possible.

13. Golgi s'est entêté à croire que les fonctions du cerveau étaient principalement prises en charge par un tissu continu dans lequel les neurones étaient enchâssés, tandis que Cajal soutenait la « doctrine neuronale ». Colauréats du prix Nobel en 1906, ils ont prononcé des discours d'acceptation antagonistes.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

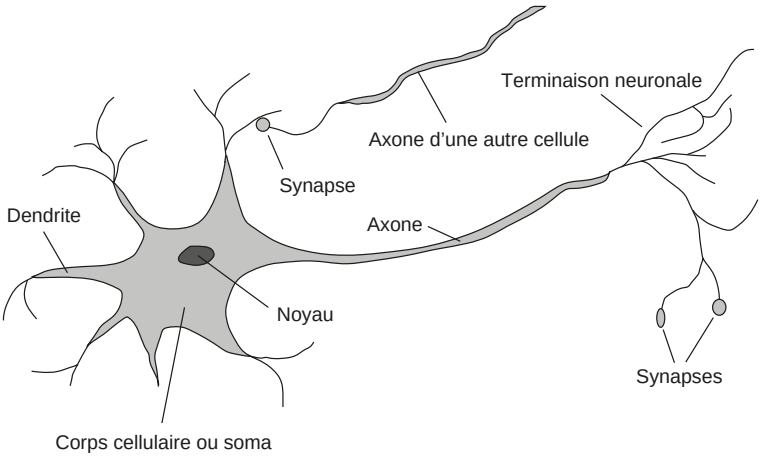

Figure 1.1 Constituants d'une cellule nerveuse ou neurone. Chaque neurone est composé d'un corps cellulaire, ou soma, qui contient un noyau. Du corps cellulaire sont issues plusieurs fibres, appelées *dendrites*, et une longue fibre, appelée *axone*. L'axone est beaucoup plus long que ne l'indique le schéma. Un axone mesure généralement 1 cm (100 fois le diamètre d'un corps cellulaire), mais il peut atteindre 1 m. Un neurone est connecté avec 10 à 100 000 autres neurones par des points de contact appelés *synapses*. Les signaux se propagent d'un neurone à l'autre par une réaction électrochimique complexe ; ils contrôlent l'activité du cerveau à court terme tout en permettant des modifications à long terme dans la connectivité des neurones. On pense que ces mécanismes forment la base de l'apprentissage dans le cerveau. L'essentiel du traitement des informations a lieu dans le cortex cérébral, l'« écorce » du cerveau. L'unité organisationnelle de base semble être une colonne de tissu d'environ 0,5 mm de diamètre, contenant environ 20 000 neurones et traversant le cortex sur toute sa profondeur (environ 4 mm chez l'homme).

La mesure de l'activité d'un cerveau intact a commencé en 1929 avec l'invention par Hans Berger (1873-1941) de l'électroencéphalogramme (EEG). Le développement de l'imagerie par résonance magnétique fonctionnelle (IRMf) [Ogawa *et al.*, 1990 ; Cabeza et Nyberg, 2001] donne aux neuroscientifiques des images d'une précision sans précédent de l'activité du cerveau, qui autorise des mesures correspondant aux processus cognitifs en cours. Ces images sont complétées par les progrès accomplis dans l'enregistrement de l'activité électrique des neurones au niveau de la cellule et par les méthodes d'**optogénétique** (Crick, 1999 ; Zemelman *et al.*, 2002 ; Han et Boyden, 2007), qui permettent la mesure et le contrôle de neurones individuels modifiés pour qu'ils soient sensibles à la lumière.

Le développement d'**interfaces neuronales directes** (Lebedev et Nicolelis, 2006) pour la perception et le contrôle moteur promet non seulement de restaurer les fonctions de personnes handicapées, mais aussi d'éclairer de nombreux aspects des systèmes nerveux. Un résultat remarquable de ces travaux est que le cerveau est capable de s'adapter pour s'interfacer avec un dispositif externe, le traitant en fait comme un autre organe ou membre sensoriel.

Le cerveau et l'ordinateur ont des propriétés assez différentes. La figure 1.2 montre que les ordinateurs ont un temps de cycle un million de fois plus rapide que le cerveau. Le cerveau compense avec beaucoup plus de stockage et d'interconnexions qu'un ordinateur personnel, même haut de gamme, bien que les plus grands supercalculateurs rivalisent avec le cerveau sur certains points.

Les futurologues se régalent avec ces nombres, en prédisant l'arrivée d'une **singularité**, un point à partir duquel les ordinateurs atteindront un niveau de performance surhumain (Vinge, 1993 ; Kurzweil, 2005 ; Doctorow et Stross, 2012), pour ensuite continuer rapidement à se perfectionner encore davantage. Cependant, les comparaisons de chiffres bruts ne sont pas particulièrement instructives. Même avec un ordinateur de capacité virtuellement infinie, nous aurions encore besoin de percées théoriques supplémentaires dans notre compréhension de l'intelligence (voir chapitre 28). Pour le dire les choses simplement, sans la bonne théorie, des machines plus rapides ne font que vous donner plus vite la mauvaise réponse.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

	Superordinateur	Ordinateur personnel	Cerveau humain
Unités de traitement	10^6 GPU + CPU	8 cœurs CPU	10^6 colonnes
	10^{15} transistors	10^{10} transistors	10^{11} neurones
Unités de stockage	10^{16} octets en RAM	10^{10} octets en RAM	10^{11} neurones
	10^{17} octets sur disque	10^{12} octets sur disque	10^{14} synapses
Durée des cycles	10^{-9} secondes	10^{-9} secondes	10^{-3} secondes
Opérations/s	10^{18}	10^{10}	10^{17}

Figure 1.2 Comparaison grossière entre un supercalculateur de pointe, Summit (Feldman, 2017), un ordinateur personnel typique de 2019 et l'humain.

1.2.5 Psychologie

- Comment pensent et agissent les hommes et les animaux ?

On fait habituellement remonter les origines de la psychologie scientifique aux travaux du physicien allemand Hermann von Helmholtz (1821-1894) et de son disciple Wilhelm Wundt (1832-1920). Helmholtz a appliqué la méthode scientifique à l'étude de la vision humaine et son *Traité d'optique physiologique* a été décrit comme « le traité le plus important sur la physique et la physiologie de la vision humaine » (Nalwa, 1993). En 1879, Wundt ouvrait le premier laboratoire de psychologie expérimentale à l'université de Leipzig. Il insistait sur les expériences soigneusement contrôlées au cours desquelles ses collaborateurs devaient réaliser une tâche perceptive ou associative tout en se livrant à une introspection de leur processus de pensée. Ces contrôles minutieux ont largement contribué à faire de la psychologie une science, mais la nature subjective des données rendait peu probable le fait que les expérimentateurs puissent un jour infirmer leurs propres théories.

À l'inverse, les biologistes qui étudient le comportement animal ne disposent pas des données fournies par introspection ; ils ont développé une méthodologie objective décrite par H. S. Jennings (1906) dans son ouvrage majeur, *Behavior of the Lower Organisms*. C'est en appliquant ce point de vue aux humains que le mouvement **béhavioriste** dirigé par John Watson (1878-1958) a rejeté *toute* théorie faisant appel à des processus mentaux en raison de l'impossibilité d'obtenir des observations fiables au moyen de l'introspection. Les béhavioristes insistent sur le fait de ne prendre en compte que les mesures objectives des percepts (ou stimuli) envoyés à un animal et les actions résultantes (ou *réponses*). Le béhaviorisme a fait de nombreuses découvertes sur les rats et sur les pigeons, mais a eu moins de succès pour ce qui est de la compréhension des humains.

La **psychologie cognitive**, qui voit le cerveau comme unité de traitement des informations, remonte au moins aux travaux de Williams James (1842-1910). Helmholtz insistait également sur l'idée que la perception nécessite une forme d'inférence logique inconsciente. Alors qu'aux États-Unis, le point de vue cognitif était grandement éclipsé par le béhaviorisme, la modélisation cognitiviste continuait à être développée à l'Unité de psychologie appliquée, dirigée par Frederic Bartlett (1886-1969) à Cambridge. Dans *The Nature of Explanation* (1943), Kenneth Craik, étudiant puis successeur de Bartlett, a rétabli avec force la légitimité de termes « mentaux » tels que « croyances » et « buts », en affirmant qu'ils étaient tout aussi scientifiques que, par exemple, les termes de « pression » et de « température » employés à propos des gaz, bien que ceux-ci soient constitués de molécules auxquelles ces propriétés ne s'appliquent pas.

Craik spécifie les trois grandes étapes d'un agent fondé sur les connaissances (*knowledge-based*) : (1) le stimulus doit être converti en représentation interne, (2) la représentation est manipulée par des processus cognitifs de manière à dériver de nouvelles représentations, (3) celles-ci sont à leur tour transformées en actions. Craik explique clairement les raisons pour lesquelles ces étapes schématisaient bien un agent (Craik, 1943) :

> Si l'organisme possède dans sa tête un « modèle réduit » de la réalité extérieure et de ses actions possibles, il est en mesure d'essayer différentes possibilités, de conclure laquelle est la meilleure, de réagir à des situations futures avant qu'elles ne surviennent, d'utiliser la connaissance des événements passés pour traiter le présent et le futur et, quelle que soit l'issue, de réagir d'une manière plus complète, plus sûre et plus compétente aux urgences auxquelles il est confronté.

Après la mort de Craik dans un accident de vélo en 1945, ses recherches ont été poursuivies par Donald Broadbent (1926-1993), dont le livre *Perception and Communication* (1958) est l'un des premiers travaux à avoir modélisé les phénomènes psychologiques comme le traitement de l'information. Dans le même temps,

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

aux États-Unis, le développement de la modélisation informatique conduisait à la création du champ des **sciences cognitives**. On peut dire de ce domaine qu'il a vu le jour au MIT, lors d'un séminaire organisé en septembre 1956, deux mois seulement après la conférence qui a vu « naître » l'IA.

Au cours de ce séminaire, George Miller a présenté *The Magic Number Seven* et Noam Chomsky, *Three Models of Language*. Allen Newell et Herbert Simon ont quant à eux exposé *The Logic Theory Machine*. Ces trois interventions importantes ont montré comment utiliser des modèles informatiques dans le cadre des problématiques respectives de la psychologie de la mémoire, du langage et de la pensée logique. Les psychologues admettent désormais généralement (bien que pas encore universellement) qu'« une théorie cognitive doit être comme un programme informatique » (Anderson, 1980), autrement dit qu'elle doit décrire le mécanisme d'une fonction cognitive en termes de traitement de l'information.

Dans le contexte de cet inventaire, nous inscrivons le domaine de l'**interaction humain-machine** (IHM) dans la rubrique psychologie. Doug Engelbart, l'un des pionniers de l'IHM, a défendu l'idée de l'**intelligence augmentée**. Il estime que les ordinateurs devraient augmenter les capacités humaines plutôt que d'automatiser les tâches humaines. En 1968, Engelbart a présenté pour la première fois, lors de la « mère de toutes les démos », une souris pour ordinateur, un système de fenêtrage, un système d'hypertexte et une vidéoconférence – tout cela dans le but de démontrer ce que les professionnels de la connaissance humaine pourraient collectivement accomplir avec une certaine augmentation de leur intelligence.

Aujourd'hui, il est fort probable que nous considérions intelligence augmentée et intelligence artificielle comme les deux faces d'une même pièce, la première mettant l'accent sur le contrôle humain et la seconde sur le comportement intelligent de la machine. Les deux sont nécessaires pour que les machines puissent être utiles aux humains.

1.2.6 Ingénierie informatique

- Comment construire un ordinateur performant ?

L'ordinateur moderne a été inventé de manière indépendante et presque simultanée par les scientifiques de trois des pays belligérants de la Seconde Guerre mondiale. L'équipe d'Alan Turing a construit en 1943 le premier ordinateur *opérationnel,* le calculateur électromécanique Heath Robinson[14], en vue d'une fonction unique : déchiffrer les messages des Allemands. La même année, cette équipe a développé Colossus, une machine puissante et polyvalente composée de tubes à vide[15]. Le Z-3, inventé par Konrad Zuse en Allemagne en 1941, a été le premier ordinateur opérationnel et *programmable*. Zuse a aussi inventé les nombres en virgule flottante et le premier langage de programmation évolué, le Plankalkül. John Atanasoff et son étudiant Clifford Berry ont assemblé le premier ordinateur *électronique*, l'ABC, entre 1940 et 1942 à l'université d'État de l'Iowa. La recherche d'Atanasoff n'a reçu que peu de soutien et de reconnaissance ; c'est l'ENIAC, développé dans le cadre d'un projet militaire secret à l'université de Pennsylvanie par une équipe incluant John Mauchly et J. Presper Eckert, qui s'est révélé le principal précurseur des ordinateurs modernes.

Depuis cette époque, chaque génération d'ordinateurs a été plus rapide, plus puissante et moins onéreuse, une tendance qui se traduit par la **loi de Moore**. Les performances ont doublé à peu près tous les 18 mois jusque vers 2005, quand les problèmes de dissipation thermique ont conduit les fondeurs à multiplier le nombre de cœurs des processeurs (CPU) plutôt qu'à augmenter la fréquence d'horloge. On s'attend actuellement à ce que les futures augmentations de fonctionnalité proviennent d'un parallélisme massif – une curieuse convergence avec les propriétés du cerveau.

Nous commençons tout juste à voir apparaître du matériel spécialement adapté aux applications d'IA, telles que le processeur graphique (GPU), le processeur tensoriel (TPU) ou le *wafer scale engine* (WSE). Des années 1960 à environ 2012, la puissance de calcul utilisée pour entraîner les meilleures applications d'apprentissage automatique a suivi la loi de Moore. À partir de 2012, les choses ont changé : de 2012 à 2018, la puissance a été multipliée par 300 000, ce qui correspond à un doublement tous les 100 jours environ (Amodei et Hernandez,

14. Une machine complexe baptisée d'après un illustrateur britannique célèbre pour ses descriptions d'appareils fantasques et absurdement compliqués destinés à des tâches routinières telles que beurrer une tartine.

15. Dans l'après-guerre, Turing a voulu utiliser ces ordinateurs pour la recherche en IA – par exemple il avait esquissé le premier programme pour jouer aux échecs (Turing *et al.*, 1953), mais le gouvernement britannique a empêché cette recherche.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

2018). Un modèle d'apprentissage automatique qui nécessitait une journée entière d'entraînement en 2014 ne prenait plus que deux minutes en 2018 (Ying *et al.*, 2018). Bien qu'elle ne soit pas encore opérationnelle, l'**informatique quantique** promet des accélérations bien plus importantes pour certaines sous-classes importantes d'algorithmes d'IA.

Bien entendu, il existait des machines à calculer bien avant l'apparition des ordinateurs électroniques. Nous avons déjà évoqué les premières machines automatisées (voir page 6). La première machine *programmable*, un métier à tisser mis au point en 1805 par Joseph-Marie Jacquard (1752-1834), utilisait des cartes perforées pour enregistrer les instructions associées au motif à réaliser.

Au milieu du XIX^e^ siècle, Charles Babbage (1792-1871) a conçu deux machines à calculer mais n'en a construit aucune. La « machine à différences » était destinée à calculer des tables mathématiques pour des projets d'ingénierie et scientifiques. Elle a finalement été construite, et a réellement fonctionné, en 1991 (Swade, 2000). Son autre projet, une « machine analytique », était beaucoup plus ambitieux : ce dispositif disposait d'une mémoire adressable, de programmes enregistrés sur la base des cartes perforées de Jacquard et de sauts conditionnels. C'est la première machine à avoir atteint le calcul universel.

Ada Lovelace, fille du poète Lord Byron et collègue de Babbage, a compris son potentiel, la décrivant comme « une machine à penser ou […] à raisonner », capable de raisonner sur « tous les sujets de l'univers » (Lovelace, 1843). Elle a également anticipé les cycles de frénésie autour de l'IA, en écrivant : « Il est souhaitable de se prémunir contre la possibilité d'idées exagérées qui pourraient surgir quant aux pouvoirs du moteur analytique. » Malheureusement, les machines de Babbage et les idées de Lovelace sont largement tombées dans l'oubli.

L'IA doit aussi beaucoup à la partie logicielle de l'informatique qui a fourni les systèmes d'exploitation, les langages de programmation et les outils nécessaires à l'écriture des programmes modernes (ainsi qu'à leur documentation). Mais il s'agit là d'un domaine pour lequel la dette a été remboursée : les travaux en IA ont fait éclore de nombreuses idées reprises en informatique. Parmi celles-ci, on peut citer le temps partagé, les interpréteurs interactifs, les ordinateurs personnels dotés d'une interface graphique et d'une souris, les environnements de développement rapide, les listes chaînées, la gestion automatique de la mémoire et les concepts clés de la programmation symbolique, fonctionnelle, déclarative et orientée objet.

1.2.7 Théorie du contrôle et cybernétique

- Comment faire en sorte que des artefacts opèrent de façon autonome ?

Ktesibios d'Alexandrie (vers 250 av. J.-C.) a construit le premier dispositif autorégulé : une horloge à eau dotée d'un régulateur pour maintenir constant le débit. Cette invention a changé la définition de ce qu'un artefact est capable de réaliser. Auparavant, seuls des êtres vivants pouvaient modifier leur comportement en réponse à des changements de leur environnement. On peut citer d'autres exemples de systèmes asservis autorégulés (*self-regulating feedback control systems*) : le régulateur du moteur à vapeur de James Watt (1736-1819) et le thermostat créé par Cornelis Drebbel (1572-1633), qui est aussi l'inventeur du sous-marin. James Clerk Maxwell (1868) a été le pionnier de la théorie mathématique du contrôle.

Norbert Wiener (1894-1964) a tenu une place centrale durant l'après-guerre dans ce qu'on appelle désormais la **théorie du contrôle** (ou **théorie de la commande**). Brillant mathématicien, Wiener a notamment travaillé avec Bertrand Russell avant de s'intéresser aux systèmes de contrôle biologiques et mécaniques et à leur rapport avec la cognition. Comme Craik (qui utilisait également des systèmes de contrôle comme modèles psychologiques), Wiener et ses collègues Arturo Rosenblueth et Julian Bigelow ont défié l'orthodoxie béhavioriste (Rosenblueth *et al.*, 1943). Selon eux, le comportement piloté par un but résultait d'un mécanisme régulateur essayant de minimiser l'« erreur » – la différence entre l'état courant (*current state*) et l'état but (*goal state*). À la fin des années 1940, Wiener, entouré de Warren McCulloch, Walter Pitts et John von Neumann, ont organisé une série de conférences fondatrices consacrées aux nouveaux modèles mathématiques et informatiques de la cognition. Le livre de Wiener, *Cybernetics* (1948), devint un best-seller qui fit prendre conscience au public des possibilités des machines artificiellement intelligentes.

Pendant ce temps, en Grande-Bretagne, W. Ross Ashby défrichait un champ similaire (Ashby, 1940). Ashby, Alan Turing, Grey Walter et d'autres ont formé le *Ratio Club* pour « ceux qui ont eu des idées analogues à

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

celles de Wiener avant que son livre ne paraisse ». L'ouvrage d'Ashby *Design for a Brain* (1948, 1952) développe l'idée qu'on pourrait créer de l'intelligence à l'aide de dispositifs **homéostatiques** contenant des boucles de rétroaction (*feedback*) appropriées permettant de garantir un comportement adaptatif stable.

La théorie du contrôle moderne, et tout particulièrement la branche appelée *contrôle stochastique optimal*, a notamment pour but la conception de systèmes minimisant une **fonction de coût** au cours du temps. Cela correspond à peu près à notre modèle standard de l'IA : la conception de systèmes au comportement optimal. Dans ce cas, pourquoi l'IA et la théorie du contrôle forment-elles deux domaines distincts, malgré les relations étroites que leurs fondateurs entretenaient ? La réponse réside dans la forte connexion qui existe entre les techniques mathématiques mises en jeu et les types de problèmes abordés par chaque domaine. Les outils de la théorie du contrôle, le calcul différentiel et l'algèbre matricielle sont plus spécifiquement destinés à des systèmes qui se décrivent sous forme d'ensembles fixes de variables continues, alors que l'IA a été fondée en partie comme un moyen d'échapper à ce qui était perçu comme des limitations de ces outils. La logique, sous sa forme informatique, a donné aux chercheurs en IA la possibilité d'étudier des problèmes comme le langage, la vision et la planification symbolique qui tombaient en dehors du champ d'investigation des puristes de la théorie du contrôle.

1.2.8 Linguistique

- Quels sont les rapports entre le langage et la pensée ?

En 1957, B. F. Skinner publiait *Verbal Behavior*. Écrit par le plus éminent expert du domaine, cet ouvrage offre un panorama complet et détaillé de l'approche béhavioriste sur le sujet de l'apprentissage du langage. Mais, curieusement, une recension du livre est devenue aussi célèbre que le livre lui-même et a eu pour conséquence de faire pratiquement disparaître tout intérêt pour le béhaviorisme. Le linguiste Noam Chomsky, son auteur, venait de publier un ouvrage exposant sa propre théorie, *Structures syntaxiques*. Il y faisait remarquer en quoi la théorie béhavioriste est impuissante à rendre compte de la notion de créativité dans le langage – elle n'explique pas comment les enfants peuvent comprendre et construire des phrases qu'ils n'ont jamais entendues auparavant. La théorie de Chomsky, fondée sur des modèles syntaxiques remontant au linguiste indien Panini (vers 350 av. J.-C.), pouvait l'expliquer et, à la différence des théories précédentes, elle était suffisamment formelle pour que sa programmation soit envisageable.

La linguistique moderne et l'IA sont donc nées à la même époque et ont évolué ensemble ; elles se croisent dans un domaine hybride appelé **linguistique computationnelle** ou **traitement automatique du langage naturel**. Le problème de la compréhension du langage s'est révélé rapidement beaucoup plus complexe qu'il n'y paraissait en 1957. Outre la connaissance de la structure des phrases, celle-ci requiert la bonne appréhension du sujet et du contexte. Cela peut sembler évident aujourd'hui, mais ce ne fut pas le cas avant les années 1960. L'essentiel des premiers travaux sur la **représentation des connaissances** (l'étude de la traduction des connaissances sous une forme permettant à l'ordinateur de raisonner) était lié au langage et nourri par des recherches en linguistique, laquelle puisait à son tour dans des décennies de recherches consacrées à l'analyse philosophique du langage.

1.3 Histoire de l'intelligence artificielle

On peut résumer les grandes étapes de l'histoire de l'IA à l'aide de la seule liste des gagnants du prix Turing : Marvin Minsky (1969) et John McCarthy (1971), pour l'élaboration des fondements du domaine sur la représentation des connaissances et le raisonnement ; Allen Newell et Herbert Simon (1975), pour les modèles symboliques de la résolution de problèmes et de la cognition humaine ; Ed Feigenbaum et Raj Reddy (1994), pour le développement de systèmes experts qui encodent la connaissance humaine permettant de résoudre des problèmes du monde réel ; Judea Pearl (2011), pour le développement de techniques de raisonnement probabiliste qui traitent l'incertitude mathématiquement ; Yoshua Bengio, Geoffrey Hinton et Yann LeCun (2019), pour avoir fait de l'apprentissage profond ou *deep learning* (réseaux de neurones multicouches) un élément essentiel de l'informatique moderne. Le reste de cette section détaille chacune des phases de l'histoire de l'IA.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

1.3.1 Genèse de l'intelligence artificielle (1943-1956)

Les premiers travaux reconnus rétrospectivement comme relevant de l'IA ont été menés par Warren McCulloch et Walter Pitts (1943). Inspirés par le travail de modélisation mathématique du professeur de Pitts, Nicolas Rashevsky (1936, 1938), ils ont puisé à trois sources : l'état du savoir sur la physiologie de base et la fonction des neurones dans le cerveau, l'analyse formelle de la logique propositionnelle de Russell et Whitehead, et la théorie du calcul de Turing. Ils ont proposé un modèle de neurones artificiels dans lequel chaque neurone est caractérisé par un état « marche » ou « arrêt », le passage à l'état « marche » se produisant en réponse à une stimulation émise par un nombre suffisant de neurones voisins. L'état d'un neurone était conçu comme « factuellement équivalent à une proposition présentant son stimulus approprié ». McCulloch et Pitts ont montré, par exemple, que toute fonction calculable peut être calculée par un réseau de neurones connectés et que tous les connecteurs logiques (ET, OU, NON, etc.) peuvent être implémentés par des structures de réseaux simples. Ils ont également suggéré que des réseaux définis de manière appropriée sont capables d'apprentissage. Donald Hebb a découvert une règle d'actualisation simple, maintenant appelée **apprentissage hebbien**, qui permet de modifier les intensités des connexions entre les neurones et qui demeure un modèle très influent aujourd'hui (Hebb, 1949).

Deux étudiants de Harvard, Marvin Minsky (1927-2016) et Dean Edmonds, ont construit le premier ordinateur à réseau de neurones en 1950, nommé SNARC. Ce système utilisait 3 000 tubes à vide et un mécanisme de pilote automatique récupéré sur un bombardier B-24 pour simuler un réseau de 40 neurones. Plus tard, à Princeton, Minsky a étudié le calcul universel sur des réseaux de neurones. Le jury de thèse de Minsky a émis des doutes quant à la nature mathématique de ce travail, mais on rapporte que von Neumann a alors déclaré : « Si ce n'est pas le cas aujourd'hui, ce le sera un jour. »

De nombreux autres exemples de travaux précurseurs peuvent être considérés comme relevant de l'IA, par exemple deux programmes de jeu de dames développés indépendamment en 1952 par Christopher Strachey à l'université de Manchester et par Arthur Samuel à IBM. Cependant, la vision d'Alan Turing a été la plus influente. Il a donné des conférences sur le sujet dès 1947 à la *London Mathematical Society* et dégagé clairement une feuille de route convaincante dans son article de 1950 « Computing Machinery and Intelligence ». C'est dans ce texte qu'il a présenté le test de Turing, l'apprentissage automatique, les algorithmes génétiques et l'apprentissage par renforcement. Il a répondu à de nombreuses objections soulevées par la possibilité même d'IA, comme décrit au chapitre 27. Il a également suggéré qu'il serait plus facile de créer une IA de niveau humain en développant des algorithmes d'apprentissage et en instruisant ensuite la machine plutôt qu'en programmant son intelligence à la main. Lors de conférences ultérieures, il a averti que la réalisation de cet objectif pourrait ne pas être la chose la plus bénéfique qui puisse arriver à l'espèce humaine.

En 1955, John McCarthy de Dartmouth College convainct Minsky, Claude Shannon et Nathaniel Rochester de l'aider à rassembler les chercheurs américains spécialisés dans la théorie des automates, les réseaux de neurones et l'étude de l'intelligence. Ils organisent alors un séminaire de deux mois à Dartmouth au cours de l'été 1956. Il y a en tout et pour tout dix participants, dont Allen Newell et Herbert Simon de Carnegie Tech[16], Trenchard More de Princeton, Arthur Samuel d'IBM, et Ray Solomonoff et Oliver Selfridge du MIT. Le projet stipule[17] :

> Nous proposons d'entreprendre une étude de l'intelligence artificielle menée par dix personnes pendant deux mois durant l'été 1956 au Dartmouth College, à Hanover, dans le New Hampshire. L'étude reposera sur la conjecture que chaque aspect de l'apprentissage ou de toute autre facette de l'intelligence peut être décrit en principe si précisément qu'une machine peut être construite pour le simuler. Nous tenterons de proposer des solutions pour que les machines puissent utiliser le langage, former des abstractions et des concepts, résoudre des types de problèmes réservés pour l'instant aux humains, et puissent se perfectionner. Nous pensons que des avancées significatives sont possibles si une équipe judicieusement sélectionnée de scientifiques travaille sur ce sujet pendant un été.

Malgré cette prévision très optimiste, le séminaire de Dartmouth n'a pas conduit à de réelles avancées. Newell et Simon ont présenté ce qui était sans doute le travail le plus abouti, un système de vérification de théorèmes

16. Maintenant Carnegie Mellon University (CMU).

17. C'était la première utilisation officielle des termes *artificial intelligence* (« intelligence artificielle ») de McCarthy. Peut-être que *computational rationality* (« rationalité computationnelle ») aurait été plus précis et moins menaçant, mais « IA » a prévalu. Au 50e anniversaire de la conférence de Dartmouth, McCarty a expliqué qu'il avait résisté aux termes *computer* ou *computational* par déférence envers Norbert Wiener, qui plaidait à l'époque pour utiliser les termes d'*analog cybernetic devices* (« dispositifs cybernétiques analogiques ») plutôt que *digital computer* pour les ordinateurs.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

mathématiques appelé *Logic Theorist* (LT). Simon a déclaré : « Nous avons inventé un programme informatique capable de penser de manière non numérique et, ce faisant, nous avons résolu le vénérable problème de la dualité du corps et de l'esprit »[18]. Peu après le séminaire, le programme était en mesure de démontrer la majorité des théorèmes du chapitre II des *Principia Mathematica* de Russell et Whitehead. Il paraît que Russell a été ravi d'apprendre que LT avait trouvé, pour un théorème, une démonstration plus courte que celle présentée dans son livre. Les éditeurs du *Journal of Symbolic Logic* ont été moins impressionnés ; ils ont rejeté un papier coécrit par Newell, Simon et *Logic Theorist*.

1.3.2 L'enthousiasme des débuts : les grandes espérances (1952-1969)

Dans les milieux intellectuels des années 1950, on préférait très largement croire qu'« une machine ne pourrait jamais faire *X* » (voir au chapitre 27 la longue liste des *X* énumérée par Turing). Les chercheurs en IA ont naturellement répondu en démontrant la faisabilité d'un *X* après l'autre. Ils se sont concentrés en particulier sur les tâches considérées comme caractéristiques de l'intelligence humaine, notamment les jeux, les puzzles, les mathématiques et les tests de QI. John McCarthy définit cette période comme celle des premiers pas : « Hé, maman… t'as vu ? Sans les mains ! »

Suite à leur succès initial avec LT, Newell et Simon ont enchaîné avec *General Problem Solver* (GPS). À la différence de LT, ce programme était conçu dès le début pour imiter la démarche des humains dans la résolution des problèmes. À l'intérieur de la classe limitée de casse-têtes qu'il était capable de traiter, il s'est avéré que l'ordre dans lequel il considérait les sous-buts et les actions possibles était comparable à celui dans lequel les humains abordaient les mêmes problèmes. C'est ainsi que GPS a certainement été le premier programme à intégrer l'approche de la « pensée humaine ». Les succès de GPS et des programmes qui ont suivi en tant que modèles de cognition ont conduit Newell et Simon (1976) à formuler la célèbre hypothèse du **système symbolique matériel**, qui énonce qu'« un système symbolique matériel contient les moyens nécessaires et suffisants pour un comportement généralement intelligent ». Ils entendaient par là que tout système faisant preuve d'intelligence doit opérer en manipulant des structures de données composées de symboles. Nous verrons que cette hypothèse a été contestée de plusieurs manières.

Chez IBM, Nathaniel Rochester et ses collègues ont produit plusieurs des premiers programmes d'IA. Herbert Gelernter (1959) a construit *Geometry Theorem Prover*, capable de démontrer des théorèmes que de nombreux étudiants trouveraient difficiles. Ce travail a été précurseur des démonstrateurs automatiques de théorèmes modernes.

De tous les travaux exploratoires réalisés durant cette période, le plus influent sur le long terme est peut-être celui d'Arthur Samuel sur les dames. En utilisant des méthodes que nous qualifierions aujourd'hui d'apprentissage par renforcement (voir chapitre 22), les programmes de Samuel apprenaient à jouer à un bon niveau amateur. Par là même, il rendait caduque l'idée que les ordinateurs n'étaient capables de faire que ce qui leur était demandé : son programme pouvait apprendre rapidement à jouer à un meilleur niveau que son créateur. Présenté à la télévision en février 1956, le programme a fait une forte impression. Comme Turing, Samuel avait du mal à obtenir du temps de calcul. Travaillant la nuit, il utilisait, dans l'usine d'IBM, des machines restées à l'étage où l'on pratiquait les tests. Le programme de Samuel est le précurseur de systèmes ultérieurs comme TD-Gammon (Tesauro, 1992), qui a été parmi les meilleurs joueurs de backgammon au monde, et AlphaGo (Silver *et al.*, 2016), qui a provoqué une stupéfaction planétaire en battant le champion du monde humain au go (voir chapitre 5).

En 1958, John McCarthy apporte deux contributions cruciales à l'intelligence artificielle. Dans le mémo n° 1 du laboratoire d'IA du MIT, il définit le langage **Lisp**, qui va devenir le langage de programmation dominant en IA pour les trente années à venir. Dans l'article intitulé « Programs with Common Sense », il formalise une conception des systèmes d'IA basés sur la connaissance et le raisonnement. Cet article décrit *Advice Taker*, un programme hypothétique qui incarnerait une connaissance générale du monde et pourrait l'utiliser pour élaborer des plans d'action. Il a illustré ce concept avec des axiomes logiques simples qui suffisent à générer un plan d'action pour se rendre à l'aéroport. Le programme était également conçu pour pouvoir accepter

18. Newell et Simon ont aussi inventé le langage de traitement de listes IPL pour écrire LT. Comme ils n'avaient pas de compilateur, ils convertissaient les instructions en code machine à la main. Pour éviter les erreurs, ils travaillaient en parallèle, chacun demandant à l'autre ses codes binaires au fur et à mesure afin de s'assurer qu'ils étaient d'accord.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

de nouveaux axiomes dans le cours normal des opérations, ce qui lui donnait la possibilité d'acquérir des compétences dans de nouveaux domaines *sans reprogrammation préalable*. *Advice Taker* incorporait donc les principes de base de la représentation des connaissances et du raisonnement, à savoir qu'il est utile de posséder une représentation formelle et explicite du monde et de ses rouages, ainsi que de pouvoir manipuler cette représentation à l'aide de processus déductifs. Il est remarquable de constater à quel point l'essentiel de cet article de 1958 demeure encore actuel. Ce papier a influencé le développement de l'IA et reste d'actualité.

C'est aussi en 1958 que Marvin Minsky s'est installé au MIT. Toutefois, sa collaboration avec McCarthy n'a guère duré. McCarthy accordait beaucoup d'importance à la logique formelle dans les représentations et les raisonnements, tandis que Minsky cherchait surtout à créer des programmes qui fonctionnent, et adoptait le cas échéant une approche antilogique. En 1963, McCarthy crée le laboratoire d'IA de Stanford. Son projet de recourir à la logique pour construire la version définitive de l'*Advice Taker* profite de la découverte réalisée en 1965 par J. A. Robinson de la méthode de résolution (un algorithme de démonstration des théorèmes de la logique du premier ordre ; voir chapitre 9). Les travaux réalisés à Stanford insistent sur l'emploi de méthodes générales de raisonnement logique. Parmi les applications de la logique prises en compte figurent les systèmes de questions-réponses et de planification de Cordell Green (1969b), ainsi que le projet de robotique Shakey du SRI (Stanford Research Institute). Ce projet présenté au chapitre 26 était le premier à intégrer complètement le raisonnement logique et l'activité physique.

Au MIT, les étudiants de Minsky avaient sélectionné des problèmes limités, dont la solution semblait exiger le recours à l'intelligence. Ces domaines limités sont, depuis, connus sous le nom de **micromondes**. Le programme SAINT (1963) de James Slagle pouvait résoudre des problèmes d'intégration en « forme close » typiques de ceux étudiés dans les classes préparatoires scientifiques. Le programme ANALOGY (1968) de Tom Evans résolvait des problèmes d'analogies géométriques tels que les tests de QI. De son côté, le programme STUDENT de Daniel Bobrow (1967) résolvait des problèmes d'algèbre élémentaire, comme :

> Si le nombre de clients obtenus par Tom est le carré de 20 % du nombre d'annonces qu'il a publiées et que le nombre de ces annonces est de 45, combien Tom a-t-il gagné de clients ?

Le plus célèbre des micromondes est le **monde des blocs**, un ensemble de blocs placés sur une table (ou plus souvent une simulation de table, voir figure 1.3). Dans ce monde, une tâche typique consiste à changer la disposition des blocs à l'aide d'un robot doté d'une pince qui peut saisir un bloc à la fois. Le monde des blocs a donné lieu au développement du projet de vision de David Huffman (1971), du travail sur la vision et la propagation des contraintes de David Waltz (1975), de la théorie de l'apprentissage de Patrick Winston (1970), du programme de compréhension du langage naturel de Terry Winograd (1972) et du planificateur de Scott Fahlman (1974).

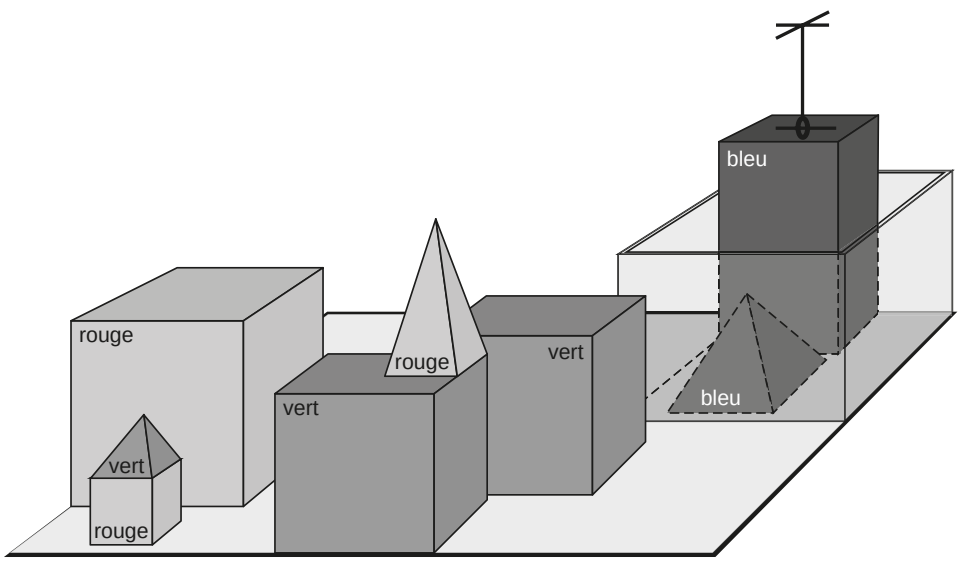

FIGURE 1.3 Scène du monde des blocs. SHRDLU (Winograd, 1972) vient de réaliser la commande « Trouve un bloc plus grand que celui que tu tiens et mets-le dans la boîte. »

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Les premiers travaux de McCulloch et de Pitts sur les réseaux de neurones ont aussi commencé à porter leurs fruits. Les études de Shmuel Winograd et de Jack Cowan (1963) ont montré qu'un grand nombre d'éléments pouvaient collectivement représenter un concept individuel, avec une augmentation correspondante de la robustesse et du parallélisme. Les méthodes d'apprentissage de Hebb ont été améliorées par Bernie Widrow (Widrow et Hoff, 1960 ; Widrow, 1962), qui a appelé ses réseaux **adalines** (pour *adaptive linear neurons*), et par Frank Rosenblatt avec ses **perceptrons** (1962). Le **théorème de convergence du perceptron** (Block *et al.*, 1962) énonce que l'algorithme d'apprentissage peut ajuster les intensités des connexions d'un perceptron afin de les corréler à n'importe quelle donnée d'entrée dès lors qu'une telle corrélation existe. Ces sujets sont traités au chapitre 20.

1.3.3 L'épreuve de la réalité (1966-1973)

Dès le début, les chercheurs en IA n'ont pas craint de faire des prédictions quant à leurs futurs succès. On cite souvent le passage suivant, écrit en 1957 par Herbert Simon :

> Mon intention n'est pas de vous surprendre ni de vous choquer, mais la manière la plus simple de résumer les faits consiste à dire qu'il existe désormais des machines capables de penser, d'apprendre et de créer. En outre, leur capacité à accomplir ces choses va rapidement s'accroître jusqu'à ce que, dans un futur proche, le champ des problèmes qu'elles pourront aborder soit coextensif à celui auquel s'applique l'esprit humain.

L'expression « dans un futur proche » est vague, mais Simon a aussi formulé des prédictions plus concrètes : dans les dix années à venir, un ordinateur serait champion d'échecs et un théorème mathématique important serait démontré par une machine. Il a fallu quarante ans au lieu de dix pour que ces prédictions se réalisent (ou presque). L'excès de confiance dont Simon a fait preuve était dû aux performances prometteuses des premiers systèmes d'IA sur des exemples simples. Cependant, dans presque tous les cas, ces premiers systèmes ont échoué sur des problèmes plus difficiles.

Deux raisons majeures expliquent cet échec. La première est que, parmi les premiers systèmes d'IA, un grand nombre était principalement basé sur une « introspection éclairée » de la manière dont les humains effectuent une tâche, plutôt que sur une analyse minutieuse de la tâche elle-même, de ce que signifie « être une solution » et de ce qu'un algorithme devrait faire pour trouver de telles solutions de manière fiable.

La deuxième raison de cet échec réside dans un manque de prise de conscience de l'impraticabilité des nombreux problèmes que l'IA essayait de résoudre. La plupart des premiers systèmes de résolution de problèmes fonctionnaient en essayant différentes combinaisons jusqu'à atteindre la solution. Cette stratégie convenait au début parce que les micromondes contenaient très peu d'objets, et donc très peu d'actions possibles et très peu de solutions. Avant le développement de la théorie de la complexité calculatoire, on s'accordait à penser qu'il suffirait de disposer de matériels plus rapides et d'une plus grande capacité de mémoire pour pouvoir aborder des problèmes de plus grande taille. L'optimisme suscité par le développement de la démonstration automatique de théorèmes par résolution a disparu sitôt que l'impossibilité de prouver des théorèmes mettant en jeu plus de quelques dizaines de faits est devenue patente. Les chercheurs venaient de comprendre que *le fait qu'un programme puisse en principe trouver une solution ne signifie pas qu'il contienne des mécanismes lui permettant de la découvrir en pratique.*

L'illusion de la puissance de calcul illimitée n'était pas propre aux seuls programmes résolvant des problèmes. Les premières expériences dans le domaine de l'**évolution artificielle** (on parle maintenant d'**algorithmes génétiques** ou de **programmation génétique** [***genetic programming***]) [Friedberg, 1958 ; Friedberg *et al.*, 1959] reposaient sur l'idée indéniablement correcte que l'application d'une série appropriée de petites mutations à un programme écrit en code machine permettrait de générer un programme caractérisé par de bonnes performances sur des tâches simples. L'idée de base consistait à essayer de manière aléatoire diverses mutations associées à un processus de sélection qui préserverait les variations les plus utiles en apparence, mais, malgré des milliers d'heures de temps machine, aucun progrès ou presque n'avait été enregistré.

Les rédacteurs du rapport Lighthill (1973) reprochaient surtout à l'IA de ne pas remédier à l'« explosion combinatoire » : c'est au vu de ce document que le gouvernement britannique a décidé de ne plus subventionner la recherche en IA que dans deux universités. (La tradition orale brosse un tableau différent, plus coloré : elle dépeint des ambitions politiques et des animosités personnelles qu'il ne saurait être question de décrire ici.)

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Une troisième difficulté est apparue en raison de certaines limitations fondamentales des structures de base utilisées pour générer un comportement intelligent. Dans leur ouvrage *Perceptrons* (1969), Minsky et Papert ont démontré par exemple que, si les perceptrons (une forme simpliste de réseaux de neurones) pouvaient apprendre tout ce qu'ils parvenaient à représenter, leur capacité de représentation restait limitée. En particulier, un perceptron à deux entrées ne pouvait pas apprendre à reconnaître les cas où ses deux entrées étaient différentes. Même si les conclusions de ces auteurs ne s'appliquent pas aux réseaux plus complexes ou multicouches, le financement des recherches sur les réseaux de neurones s'est raréfié au point de se réduire à presque rien. Ironie du sort, les nouveaux algorithmes d'apprentissage par rétropropagation qui allaient provoquer une importante résurgence de la recherche sur les réseaux de neurones à la fin des années 1980 et à nouveau dans les années 2010 avaient déjà été développés dans d'autres contextes au début des années 1960 (Kelley, 1960; Bryson, 1962).

1.3.4 Systèmes experts (1969-1986)

Le paradigme de résolution de problèmes élaboré au cours de la première décennie de recherche en IA consistait en un mécanisme d'exploration d'ordre général qui essayait d'enchaîner des étapes de raisonnement élémentaires pour trouver des solutions complètes. De telles approches ont été qualifiées de **méthodes faibles** car, quoique générales, elles ne supportent pas le changement d'échelle pour résoudre des problèmes plus grands ou plus difficiles. L'alternative aux méthodes faibles est de recourir à des connaissances plus puissantes et spécifiques au domaine concerné, qui permettent de considérer des étapes de raisonnement plus importantes et de gérer plus facilement les cas typiques rencontrés dans des domaines d'expertise limités. On pourrait dire que, pour résoudre un problème difficile, il est presque obligatoire d'en connaître la solution à l'avance.

Le programme DENDRAL (Buchanan *et al.*, 1969) constitue l'un des premiers exemples de cette approche. Il a été développé à Stanford, où Ed Feigenbaum (ancien étudiant de Herbert Simon), Bruce Buchanan (philosophe devenu chercheur en informatique) et Joshua Lederberg (généticien lauréat du prix Nobel) ont conjointement résolu le problème de l'inférence d'une structure moléculaire à partir des informations fournies par un spectromètre de masse. L'entrée du programme est composée de la formule élémentaire d'une molécule (par exemple, $C_6H_{13}NO_2$) et du spectre de masse des différents fragments de la molécule générés lorsqu'elle est bombardée par un faisceau d'électrons. Par exemple, un spectre de masse peut contenir un pic en $m = 15$, soit la masse d'un groupe méthyle (CH_3).

La version naïve du programme générait toutes les structures composables à partir de la formule de départ et prédisait ensuite le spectre de masse observable pour chacune, après quoi il comparait cette donnée au spectre effectif. Comme on pouvait s'y attendre, il est apparu que cette procédure était impraticable même pour des molécules de taille modérée. Les concepteurs de DENDRAL ont alors consulté des experts en chimie analytique, qui leur ont expliqué qu'ils procédaient en recherchant dans les spectres le profil de pics bien connus, suggérant l'existence de sous-structures classiques dans la molécule. Par exemple, la règle suivante sert à reconnaître un groupement carbonyle (C=O, dont le poids est 28).

> **si** M est la masse totale de la molécule et il y a deux pics à x_1 et x_2 tels que :
> (a) $x_1 + x_2 = M + 28$; (b) $x_1 - 28$ est un pic élevé; (c) $x_2 - 28$ est un pic élevé;
> (d) au moins une des deux valeurs x_1 et x_2 est élevée,
> **alors** il y a un groupement carbonyle.

Le fait de détecter la présence d'une sous-structure particulière dans une molécule réduit considérablement le nombre de candidats possibles. Selon ses concepteurs, la puissance de la nouvelle version de DENDRAL provenait de ce qu'il incorporait la connaissance utile en spectrographie de masse non sous la forme de principes premiers, mais sous la forme de « recettes de cuisine » efficaces (Feigenbaum *et al.*, 1971). L'intérêt de DENDRAL est qu'il est le premier système à avoir réussi à faire un usage *intensif* des *connaissances* : son expertise découlait d'un grand nombre de règles spécialisées. En 1971, Feigenbaum et ses collègues de Stanford ont lancé le projet HPP (*Heuristic Programming Project*) afin d'étudier dans quelle mesure la nouvelle méthodologie des **systèmes experts** pouvait s'étendre à d'autres domaines.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Le système MYCIN, pour le diagnostic des infections du sang, a été la réalisation majeure suivante. Avec 450 règles environ, MYCIN présentait les mêmes performances que certains experts et obtenait des résultats bien meilleurs que des médecins fraîchement diplômés. Ce système différait de DENDRAL sur deux points importants. En premier lieu, il n'existait pas de modèle théorique duquel on aurait pu extraire les règles de MYCIN : il a fallu les acquérir par de longs entretiens avec des experts. En second lieu, les règles devaient refléter l'incertitude inhérente au savoir médical : MYCIN intégrait un calcul de l'incertitude fondé sur des **facteurs de certitude** (voir chapitre 13), qui paraissait (à l'époque) bien correspondre à la manière dont les médecins évaluent l'impact des résultats d'analyse médicale sur le diagnostic.

Le premier système expert commercial couronné de succès, R1, a été mis en service chez Digital Equipment Corporation (McDermott, 1982). Le programme a aidé à configurer les commandes de nouveaux systèmes informatiques ; en 1986, il permettait à la société d'économiser environ 40 millions de dollars par an. En 1988, la division IA de DEC avait déployé 40 systèmes experts, et d'autres étaient en préparation. DuPont avait 100 systèmes en service et 500 en cours de développement. Presque toutes les grandes entreprises américaines avaient leur propre division IA et utilisaient ou étudiaient les systèmes experts.

L'importance de la connaissance spécifique du domaine considéré est également devenue manifeste dans le champ de la compréhension du langage naturel. Malgré le succès du système SHRDLU de Winograd, ses méthodes ne s'étendaient pas à des tâches plus générales : pour des problèmes tels que la résolution d'ambiguïtés, il utilisait des règles simples qui s'appuyaient sur la portée limitée du monde des blocs.

Selon certains chercheurs, comme Eugene Charniak au MIT et Roger Schank à Yale, toute compréhension substantielle du langage suppose une connaissance profonde du monde et une méthode générique d'utilisation de cette connaissance. Schank alla même plus loin en déclarant que « la syntaxe n'existe pas », ce qui provoqua un tollé chez de nombreux linguistes mais permit de lancer une discussion fructueuse. Schank et ses étudiants construisirent une série de programmes (Schank et Abelson, 1977 ; Wilensky, 1978 ; Schank et Riesbeck, 1981 ; Dyer, 1983) qui avaient tous pour tâche la compréhension du langage naturel. Cependant, l'accent portait moins sur le langage *en soi* que sur les problèmes de représentation des connaissances et de raisonnement sur les connaissances nécessaires à la compréhension du langage.

L'apparition généralisée d'applications se rapportant à des problèmes du monde réel a conduit au développement d'outils de représentation des connaissances et de raisonnement. Certains étaient fondés sur la logique : c'est le cas du langage Prolog, très apprécié en Europe et au Japon, et de la famille de langages PLANNER, plus répandue aux États-Unis. D'autres, fondés sur le concept des **schémas** (***frames***) développé par Minsky (1975), ont adopté une approche plus structurée : ils assemblent des observations relatives à des types particuliers d'objets et d'événements, puis ordonnent ces types dans une vaste hiérarchie taxonomique analogue à celle de la biologie.

En 1981, le gouvernement japonais lançait le projet « Cinquième génération », plan décennal qui prévoyait de construire des ordinateurs massivement parallèles, intelligents, programmés en Prolog. Le budget annoncé dépassait les 1,3 milliard de dollars en valeur actualisée. En réponse à cette annonce, les États-Unis ont créé le consortium de recherche MCC (Microelectronics and Computer Technology Corporation), qui visait à préserver la compétitivité américaine en ce domaine. Dans les deux cas, l'IA faisait partie d'un effort plus vaste, comprenant des recherches sur les architectures de puces et sur les interfaces homme-machine. En Grande-Bretagne, le rapport Alvey rétablit les financements supprimés à la suite du rapport Lighthill. Cependant, aucun de ces projets n'a jamais atteint ses objectifs ambitieux en termes de nouvelles capacités d'IA ou d'impact économique.

Dans l'ensemble, l'industrie de l'IA connaît alors un développement intense, son chiffre d'affaires passant de quelques millions de dollars en 1980 à plusieurs milliards de dollars en 1988, avec des centaines d'entreprises vendant des systèmes experts, des systèmes de vision par ordinateur, des robots, ainsi que du matériel et du logiciel spécialisé pour ces applications.

Vient peu après une période de stagnation, dite « hiver de l'IA », au cours de laquelle de nombreuses entreprises se retrouvent sur le carreau parce qu'elles n'ont pas tenu leurs promesses extravagantes. Il s'avère difficile de construire et de maintenir des systèmes experts pour des domaines complexes, en partie parce que les méthodes de raisonnement utilisées par les systèmes s'effondrent face à l'incertitude, en partie parce que les systèmes ne sont pas capables d'apprendre de l'expérience.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

1.3.5 Retour des réseaux de neurones (de 1986 à nos jours)

Au milieu des années 1980, au moins quatre groupes de chercheurs différents réinventent l'algorithme d'apprentissage par **rétropropagation**, initialement mis au point par Bryson et Ho en 1969. Cet algorithme est appliqué à de nombreux problèmes d'apprentissage en informatique et en psychologie, et la publication des résultats dans la collection *Parallel Distributed* (Rumelhart et McClelland, 1986) suscite énormément d'enthousiasme.

Certains estimaient alors que les modèles dits **connexionnistes** concurrenceraient directement tant les modèles symboliques prônés par Newell et Simon que l'approche logiciste de McCarthy et d'autres chercheurs. Il peut sembler évident que les humains manipulent des symboles à un certain niveau – l'ouvrage de l'anthropologue Terrence Deacon, *The Symbolic Species* (1997), va même jusqu'à préciser qu'il s'agirait là d'un *trait distinctif* du genre humain. À l'opposé, Geoff Hinton, figure de proue de la résurgence des réseaux de neurones dans les années 1980 et 2010, décrit les symboles comme l'« éther de l'IA », en référence au médium imaginaire par lequel de nombreux physiciens du XIX^e^ siècle croyaient que les ondes électromagnétiques se propageaient. Certes, à y regarder de plus près, de nombreux concepts qu'on nomme par le langage ne sont pas définissables sous forme logique de conditions nécessaires et suffisantes, ce qui réduit à néant les espoirs qu'avaient les premiers chercheurs en IA de les représenter sous forme axiomatique. Il se peut que les modèles connexionnistes forment des concepts internes d'une manière plus fluide et plus imprécise qui convient mieux au désordre du monde réel. Ils ont également la capacité d'apprendre à partir d'exemples : pour un problème donné, ils peuvent comparer la valeur de sortie prédite à la valeur observée et modifier leurs paramètres pour diminuer la différence, ce qui les rend plus susceptibles d'obtenir de meilleurs résultats avec de futurs exemples.

1.3.6 Raisonnement probabiliste et apprentissage automatique (de 1987 à nos jours)

La fragilité des systèmes experts a conduit à une nouvelle approche, plus scientifique, intégrant les probabilités plutôt que la logique booléenne, l'apprentissage automatique (*machine learning*) plutôt que le codage manuel, et les résultats expérimentaux plutôt que les revendications philosophiques[19]. Il est devenu plus fréquent de reprendre des théories existantes que d'en proposer de nouvelles, de fonder des propositions sur des théorèmes rigoureux ou une solide méthodologie expérimentale (Cohen, 1995) plutôt que sur l'intuition et de préférer les applications du monde réel aux exemples de laboratoire.

Les banques de problèmes de référence (*benchmarks*) sont devenues monnaie courante pour démontrer les progrès réalisés. On peut citer celles de l'université de Californie à Irvine pour les jeux de données destinés à l'apprentissage automatique, la compétition IPC (International Planning Competition) pour les algorithmes de planification, le corpus LibriSpeech pour la reconnaissance de la parole, le jeu de données MNIST pour la reconnaissance de chiffres manuscrits, les jeux de données ImageNet et COCO pour la reconnaissance d'objets sur images, SQuAD pour les réponses à des questions en langage naturel, le concours WMT pour la traduction automatique et les compétitions internationales SAT pour les solveurs de satisfaisabilité booléens.

L'IA a en partie été fondée dans un esprit de rébellion face aux limites des domaines existants, comme la théorie du contrôle et les statistiques, mais cela ne l'a pas empêchée à l'époque de faire siens les résultats positifs de ces disciplines. Comme David McAllester (1998) l'a écrit :

> Au cours de la première période de l'IA, tout portait à croire que de nouvelles formes de calcul symbolique, par exemple les schémas et les réseaux sémantiques, allaient largement contribuer à l'obsolescence des théories classiques. Cela déboucha sur une sorte d'isolationnisme qui sépara clairement l'IA du reste de l'informatique : c'est cet isolationnisme qui est actuellement abandonné. On reconnaît que l'apprentissage automatique ne doit pas être isolé de la théorie de l'information, que le raisonnement en environnement incertain ne doit pas être dissocié des modélisations stochastiques, que l'exploration ne doit pas être séparée de l'optimisation et du contrôle classiques, et que le raisonnement automatique ne doit pas être disjoint des méthodes formelles et de l'analyse statique.

19. Certains y ont vu une victoire des ***neats*** (ceux qui pensent que les théories de l'IA doivent être mathématiquement rigoureuses) sur les ***scruffies*** (ceux qui, tenant à essayer toutes sortes d'idées, préfèrent écrire des programmes puis voir ce qui semble marcher) – on pourrait dire les cigales et les fourmis –, mais ces deux approches sont aussi importantes l'une que l'autre. Un virage en direction de la « propreté » indique que le domaine est devenu stable et mature. L'accent mis actuellement sur l'apprentissage profond peut représenter une résurgence des *scruffies*.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Le domaine de la reconnaissance de la parole illustre ce schéma. Dans les années 1970, une grande diversité d'architectures et d'approches a été tentée. Nombre d'entre elles n'étaient que des constructions *ad hoc* et fragiles, et ne fonctionnaient que sur quelques exemples soigneusement sélectionnés. Dans les années 1980, des approches utilisant les **modèles de Markov cachés (MMC)**, ou ***hidden Markov models***, ont dominé cette discipline, et deux caractéristiques des MMC l'expliquent. Premièrement, ils sont étayés par une théorie mathématique rigoureuse, ce qui a permis aux chercheurs dans le domaine de la reconnaissance de la parole de s'appuyer sur plusieurs décennies de résultats mathématiques acquis dans d'autres domaines. Deuxièmement, ils sont générés par un processus d'apprentissage mené sur un corpus important de données vocales réelles, ce qui assure la robustesse des performances : sur des tests rigoureux effectués en aveugle, les modèles MMC ont constamment amélioré leurs scores. En conséquence, la technologie de reconnaissance de la parole et le domaine voisin de reconnaissance de caractères manuscrits ont migré vers des applications industrielles et grand public très répandues. Notez qu'aucun scientifique n'avance que les humains utilisent des MMC pour la reconnaissance de la parole ; l'idée est que les MMC fournissent un cadre mathématique intéressant pour comprendre et résoudre le problème. Nous verrons cependant en section 1.3.8 que l'apprentissage profond a quelque peu bouleversé ce récit si confortable.

1988 a été une année importante pour les liens entre l'IA et d'autres domaines, notamment les statistiques, la recherche opérationnelle, la théorie de la décision et la théorie du contrôle. L'ouvrage de Judea Pearl intitulé *Probabilistic Reasoning in Intelligent Systems* (1988) a conduit à une nouvelle acceptation de la théorie des probabilités et de la décision en IA. Le développement des **réseaux bayésiens** par Pearl a débouché sur un formalisme rigoureux et efficace pour représenter les connaissances incertaines ainsi que sur des algorithmes pratiques pour le raisonnement probabiliste. Les chapitres 12 à 16 traitent ce point, ainsi que les développements plus récents, qui ont considérablement augmenté le pouvoir expressif des formalismes probabilistes. Le chapitre 20 décrit des méthodes d'apprentissage pour les réseaux bayésiens et les modèles connexes à partir de données.

Les travaux de Rich Sutton sont une deuxième contribution majeure de l'année 1988. Il a connecté l'apprentissage par renforcement, utilisé dans le programme de jeu de dames d'Arthur Samuel dans les années 1950, à la théorie des processus décisionnels de Markov (PDM), développée dans le domaine de la recherche opérationnelle. Un grand nombre de travaux ont suivi, reliant la recherche sur la planification en IA aux PDM, et le domaine de l'apprentissage par renforcement a trouvé des applications en robotique et en contrôle des processus, tout en acquérant de solides bases théoriques.

L'une des conséquences de la toute nouvelle reconnaissance de l'IA pour les données, la modélisation statistique, l'optimisation et l'apprentissage automatique a été la réintégration progressive de sous-domaines tels que la vision par ordinateur, la robotique, la reconnaissance de la parole, les systèmes multiagents, le traitement du langage naturel, qui s'étaient quelque peu séparés de l'IA de base. Ce processus de réunification a engendré des bénéfices considérables tant en termes d'applications – par exemple, le déploiement de robots pratiques s'est considérablement développé au cours de cette période – qu'en termes de meilleure compréhension théorique des problèmes fondamentaux de l'IA.

1.3.7 Les *big data* (de 2001 à nos jours)

Les progrès remarquables de la puissance de calcul et la création du World Wide Web ont facilité la constitution de très grands jeux de données – ce qu'on appelle parfois les ***big data*** (« mégadonnées »). Ces jeux de données comprennent des milliards de mots de texte, des milliards d'images et des milliards d'heures d'audio et de vidéo, ainsi que de vastes quantités de données génomiques, de suivi de véhicules, de parcours de navigation, de réseaux sociaux, etc.

Cela a conduit à la mise au point d'algorithmes d'apprentissage spécialement conçus pour tirer profit de très grands jeux de données. Souvent, la grande majorité des exemples de ces jeux de données ne sont pas *étiquetés*. Par exemple, dans les travaux déterminants de Yarowsky sur la désambiguïsation du sens des mots (1995), les occurrences d'un mot tel que *plant* (« plante » ou « usine ») ne sont pas étiquetées dans le jeu des données pour indiquer si elles se réfèrent à la flore ou à une usine. Cependant, avec des jeux de données suffisamment importants, des algorithmes d'apprentissage appropriés peuvent atteindre une précision de plus de 96 % sur la tâche qui consiste à identifier le sens qui était souhaité dans une phrase. En outre, Banko et Brill (2001) ont fait valoir que l'amélioration des performances obtenue par l'augmentation de la taille du jeu de données

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

de deux ou trois ordres de grandeur l'emporte sur toute autre amélioration pouvant être obtenue en ajustant l'algorithme.

Un phénomène analogue semble se produire dans le domaine de la vision par ordinateur. Par exemple, Hays et Efros (2007) ont mis au point une méthode intelligente pour résoudre le problème du remplissage des vides dans une photographie (ces vides étant causés soit par des dommages, soit par l'effacement d'anciens amis), en mélangeant les pixels d'images similaires. Ils ont constaté que la technique fonctionne mal avec une base de données de seulement quelques milliers d'images, mais franchit un seuil de qualité avec des millions d'images. Peu après, la disponibilité de dizaines de millions d'images dans la base de données ImageNet (Deng *et al.*, 2009) a déclenché une révolution dans le domaine de la vision par ordinateur.

La disponibilité de données massives et le passage à l'apprentissage automatique ont permis à l'IA de retrouver son attrait commercial (Havenstein, 2005 ; Halevy *et al.*, 2009). En 2011, la victoire du système Watson d'IBM face à des champions humains au jeu télévisé Jeopardy! est un événement qui a eu un impact majeur sur la perception de l'IA par le public. Les *big data* ont constitué un facteur déterminant dans cette victoire. Cependant, ce n'est qu'en 2011 que les méthodes d'apprentissage profond ont réellement pris leur essor. Cela s'est produit d'abord avec la reconnaissance de la parole, puis avec la reconnaissance visuelle d'objets.

1.3.8 Apprentissage profond (de 2011 à nos jours)

Le terme **apprentissage profond** (***deep learning***) désigne l'apprentissage automatique utilisant plusieurs couches d'unité de calcul simples et ajustables. De tels réseaux ont fait l'objet d'expériences dès les années 1970, et ils ont rencontré un certain succès dans la reconnaissance de chiffres manuscrits dans les années 1990 (LeCun *et al.*, 1995) sous la forme de **réseaux de neurones convolutifs**. Cependant, ce n'est qu'en 2011 que les méthodes d'apprentissage profond ont vraiment décollé. Cela s'est produit d'abord en reconnaissance de la parole, puis avec l'identification visuelle des objets.

Lors de la compétition ImageNet de 2012 dans laquelle on demandait de classer des images, en attribuant à chaque image une seule catégorie parmi un millier de possibilités (tatou, bibliothèque, tire-bouchon, etc.), un système avec apprentissage profond créé par l'équipe de Geoffrey Hinton à l'université de Toronto (Krizhevsky *et al.*, 2013) a montré une amélioration spectaculaire par rapport aux systèmes précédents, qui étaient basés sur l'utilisation d'attributs prédéfinis. Depuis lors, les systèmes d'apprentissage profond ont dépassé les performances humaines sur certaines tâches de vision (et sont à la traîne pour d'autres). Des gains similaires ont été observés pour la reconnaissance de la parole, la traduction automatique, le diagnostic médical et les jeux. L'utilisation d'un réseau profond pour représenter la fonction d'évaluation a contribué aux victoires de AlphaGo sur les meilleurs joueurs humains de go (Silver *et al.*, 2016, 2017, 2018).

Ces succès remarquables ont suscité un regain d'intérêt pour l'IA parmi les étudiants, les entreprises, les investisseurs, les gouvernements, les médias et le grand public. Chaque semaine semble voir paraître l'annonce d'une nouvelle application d'IA approchant ou dépassant les performances humaines, souvent accompagnée de son lot de spéculations quant à l'accélération des progrès ou au contraire l'arrivée d'un nouvel hiver de l'IA.

L'apprentissage profond repose en grande partie sur du matériel performant. Quand l'unité centrale d'un ordinateur classique effectue 10^9 ou 10^{10} opérations par seconde, un algorithme d'apprentissage profond fonctionnant sur du matériel spécialisé (par exemple, GPU, TPU ou FPGA) peut effectuer entre 10^{14} et 10^{17} opérations par seconde, principalement sous la forme d'opérations matricielles et vectorielles hautement parallélisées. Bien sûr, l'apprentissage profond dépend aussi de la disponibilité de grandes quantités de données d'entraînement, et de quelques astuces algorithmiques (voir chapitre 21).

1.4 État de l'art

L'initiative de l'université de Stanford *One Hundred Year Study on AI* (« IA : une vision à 100 ans », également connue sous le nom d'AI100) réunit des groupes d'experts pour établir des rapports sur l'état de l'art en matière d'IA. Le rapport de 2016 (Stone *et al.*, 2016 ; voir aussi Grosz et Stone, 2018) conclut que « l'on peut s'attendre à une augmentation substantielle des utilisations futures des applications de l'IA, en particulier un plus grand nombre de voitures autonomes, de diagnostics médicaux et de traitements ciblés, ainsi qu'une assistance physique pour les soins aux personnes âgées ». Le rapport signale aussi que « la société se trouve main-

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

tenant à un moment crucial pour déterminer comment déployer les technologies basées sur l'IA de manière à promouvoir plutôt qu'à entraver les valeurs démocratiques telles que la liberté, l'égalité et la transparence ». L'initiative AI100 produit également un **index IA**, qu'on trouve sur `aiindex.org`, permettant de suivre les progrès accomplis. Voici quelques faits marquants des rapports de 2018 et 2019 (l'année de référence utilisée pour les comparaisons est l'an 2000, sauf indication contraire) :

- **Publications.** Le nombre de publications sur l'IA a été multiplié par 20 entre 2010 et 2019, pour atteindre environ 20 000 par an. La catégorie la plus représentée est celle de l'apprentissage automatique. Sur `arXiv.org`, les articles concernant ce sujet ont doublé chaque année entre 2009 et 2017. La vision par ordinateur et le traitement du langage naturel ont occupé la deuxième place des thèmes les plus abordés.
- **Ressenti.** Environ 70 % des articles de presse sur l'IA sont neutres, mais les articles positifs sont passés de 12 % en 2016 à 30 % en 2018. Les questions les plus courantes sont d'ordre éthique : la confidentialité des données et les biais algorithmiques.
- **Étudiants.** Les inscriptions aux cours ont été multipliées par 5 aux États-Unis et par 16 au niveau international par rapport à 2010. L'IA est la spécialisation la plus populaire en informatique.
- **Égalité.** Les professeurs d'IA sont environ à 80 % des hommes et à 20 % des femmes dans le monde. Les chiffres sont similaires pour les étudiants en doctorat et les personnes en poste dans l'industrie.
- **Conférences.** La participation à NeurIPS a augmenté de 800 % depuis 2012 pour atteindre 13 500 participants. Les autres conférences connaissent une croissance annuelle d'environ 30 %.
- **Industrie.** Le nombre de créations d'entreprise en IA a été multiplié par 20 aux États-Unis, pour atteindre plus de 800.
- **Internationalisation.** La Chine publie plus d'articles par an que les États-Unis et à peu près autant que l'ensemble de l'Europe. Toutefois, les auteurs américains récoltent 50 % de citations en plus que les auteurs chinois. Singapour, le Brésil, l'Australie, le Canada et l'Inde sont les pays qui connaissent la croissance la plus rapide en termes de nombre d'embauches en IA.
- **Vision.** Le taux d'erreur dans la détection d'objets (tel qu'il a été atteint dans le cadre du challenge LSVRC (Large-Scale Visual Recognition Challenge) a chuté de 28 % en 2010 à 2 % en 2017, surpassant les performances humaines. La précision des réponses visuelles aux questions ouvertes (VQA) s'est améliorée de 55 % à 68 % depuis 2015, mais reste inférieure à la performance humaine, qui atteint 83 %.
- **Rapidité.** Le temps d'entraînement nécessaire pour la reconnaissance d'images a été divisé par 100 au cours des deux dernières années. La puissance de calcul utilisée dans les principales applications d'IA double tous les 3,4 mois.
- **Langage.** La précision sur la réponse aux questions, mesurée par le score F1 sur le jeu de données de questions-réponses de Stanford (SQuAD), est passée de 60 à 95 de 2015 à 2019 ; sur la variante SQuAD 2, les progrès ont été encore plus rapides, passant de 62 à 90 en un an seulement. Ces deux scores dépassent les performances humaines.
- **Critères humains.** En 2019, les systèmes d'IA ont atteint ou dépassé les performances humaines aux échecs, au go, au poker, au Pac-Man, à Jeopardy!, à la détection d'objets sur ImageNet, en reconnaissance de la parole, dans un domaine limité, en traduction chinois-anglais, également dans un champ restreint, à Quake III, à Dota 2, à StarCraft II, à divers jeux Atari, à la détection du cancer de la peau, à la détection du cancer de la prostate, en repliement de protéines et en diagnostic de la rétinopathie diabétique.

Si cela doit arriver, quand les systèmes d'IA atteindront-ils un niveau de performance humain sur une grande diversité de tâches ? Les experts en IA interrogés par Ford (2018) proposent un large éventail d'années cibles, de 2029 à 2200, avec une moyenne de 2099. Dans une enquête similaire (Grace *et al.*, 2017), 50 % des personnes interrogées estiment que cela pourrait se produire d'ici 2066, 10 % qu'on peut s'y attendre dès 2025, et quelques-unes ont déclaré « jamais ». Les experts sont également divisés sur la question de savoir si ce but nécessite de nouvelles percées technologiques, ou seulement un perfectionnement des approches actuelles. Mais ne prenez pas leurs prévisions trop au sérieux ; comme le démontre Philip Tetlock (2017), dans le domaine de la prospective, les experts ne sont pas meilleurs que les amateurs.

Comment les futurs systèmes d'IA fonctionneront-ils ? Nous ne pouvons pas encore le dire. Comme nous venons de le voir dans cette section, le domaine de l'IA a vu son paradigme dominant évoluer au cours du

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

temps : tout d'abord l'idée audacieuse que l'intelligence était atteignable par une machine, puis qu'elle pouvait être réalisée en encodant les connaissances d'experts grâce à la logique, ensuite que les modèles probabilistes du monde seraient l'outil principal, et plus récemment que l'apprentissage automatique induirait des modèles qui pourraient ne reposer sur aucune théorie bien comprise, quelle qu'elle soit. L'avenir nous dira quel sera le prochain modèle.

Que peut faire l'IA aujourd'hui ? Peut-être pas autant que certains articles médiatiques plus optimistes voudraient le laisser croire, mais tout de même beaucoup. Voici quelques exemples.

Véhicules robotisés. L'histoire des véhicules robotisés remonte aux voitures radiocommandées des années 1920, mais les premières démonstrations de conduite autonome sur route sans guidage spécifique ont eu lieu dans les années 1980 (Kanade *et al.*, 1986 ; Dickmanns et Zapp, 1987). Après des démonstrations réussies de conduite sur des routes de terre dans le cadre du 132-mile DARPA Grand Challenge en 2005 (Thrun, 2006) et sur des rues avec circulation dans le cadre du 2007 Urban Challenge, la course au développement des voitures autonomes a sérieusement commencé. En 2018, les véhicules d'essai Waymo ont franchi le cap des 10 millions de miles (16 millions de kilomètres) parcourus sur la voie publique sans accident grave, le conducteur humain n'intervenant pour prendre le contrôle qu'une fois tous les 10 000 kilomètres. Peu après, l'entreprise a commencé à offrir un service commercial de taxis robotisés.

Dans les airs, le Rwanda utilise depuis 2016 des drones autonomes à voilure fixe pour livrer du sang à travers le pays. Les quadricoptères effectuent des manœuvres acrobatiques spectaculaires, explorent des bâtiments pour en construire des cartes en 3D et s'assemblent tout seuls en formations autonomes.

Locomotion. BigDog, un robot quadrupède développé par Raibert *et al.* (2008), a bouleversé nos *a priori* sur la façon dont les robots se déplacent : ce n'est plus la démarche lente, raide, latérale des robots des films hollywoodiens, mais quelque chose ressemblant réellement à un animal et capable de se rattraper lorsqu'on le pousse ou lorsqu'il glisse sur une plaque de glace. Atlas, un robot humanoïde, non seulement marche sur un terrain accidenté mais saute sur des caisses et exécute des sauts périlleux (Ackerman et Guizzo, 2016).

Planification et programmation autonomes. À 100 millions de kilomètres de la Terre, le programme Remote Agent de la NASA est devenu le premier planificateur autonome embarqué pour contrôler le séquencement des opérations d'un vaisseau spatial (Jonsson *et al.*, 2000). Remote Agent a généré des plans à partir d'objectifs de haut niveau spécifiés depuis le sol et a surveillé l'exécution de ces plans – en détectant, en diagnostiquant et en résolvant les problèmes au fur et à mesure qu'ils se présentaient. Aujourd'hui, le système de planification Europa (Barreiro *et al.*, 2012) est utilisé pour les opérations quotidiennes des rovers de la NASA sur Mars, et le système Sextant (Winternitz, 2017) permet une navigation autonome dans l'espace profond, au-delà du système GPS global.

Au cours de la guerre du Golfe survenue en 1991, les forces armées des États-Unis ont déployé Dart (*Dynamic Analysis and Replanning Tool*), un outil d'analyse et de replanification dynamique (Cross et Walker, 1994) qui automatise la gestion de la logistique et de la programmation des transports. Cela impliquait jusqu'à 50 000 véhicules, bateaux et soldats à la fois, et il fallait également tenir compte des points de départ, des destinations, des routes, des moyens de transport disponibles, des capacités des ports et des aéroports et de la résolution des conflits, entre les nombreux paramètres. Selon la DARPA (Defense Advanced Research Project Agency), cette application a remboursé à elle seule les sommes investies depuis 30 ans dans l'IA.

Tous les jours, des services d'appel de voitures avec chauffeur comme Uber et de cartographie comme Google Maps fournissent rapidement des itinéraires optimaux à des centaines de millions d'utilisateurs, en tenant compte des conditions de circulation en cours et futures.

Traduction automatique. Les systèmes de traduction automatique en ligne arrivent désormais à lire des documents dans plus de 100 langues, représentant les langues maternelles de plus de 99 % de la population mondiale, et restituent des centaines de milliards de mots par jour pour des centaines de millions d'utilisateurs. Bien que les phrases produites ne soient pas parfaites, elles sont généralement suffisantes pour la compréhension. Pour les langues proches comme le français et l'anglais, pour lesquelles on dispose de beaucoup de données d'entraînement, les traductions dans un domaine restreint sont voisines du niveau humain (Wu, Schuster *et al.*, 2016).

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Reconnaissance de la parole. En 2017, Microsoft a montré que son logiciel *Conversational Speech Recognition System* avait atteint un taux d'erreur de 5,1 % sur les mots, similaire à la performance humaine sur la tâche *Switchboard*, qui consiste à transcrire des conversations téléphoniques (Xiong *et al.*, 2017). Environ un tiers des interactions homme-machine dans le monde se font désormais par la voix plutôt que par le clavier ; Skype fournit une traduction vocale en temps réel dans 10 langues. Alexa, Siri, Cortana et Google proposent des assistants qui peuvent répondre aux questions et effectuer des tâches pour l'utilisateur ; par exemple le service Google Duplex utilise la reconnaissance de la parole et la synthèse vocale pour effectuer des réservations au restaurant pour les utilisateurs, en menant une conversation fluide en leur nom.

Recommandation. Des sociétés telles qu'Amazon, Facebook, Netflix, Spotify, YouTube, Walmart, etc., utilisent l'apprentissage automatique pour recommander ce qui pourrait vous plaire en fonction de vos expériences passées et de celles des personnes ayant votre profil. Le domaine des systèmes de recommandation a une longue histoire (Resnick et Varian, 1997) mais évolue rapidement en raison des nouvelles méthodes d'apprentissage profond qui analysent le contenu (texte, musique, vidéo) ainsi que les historiques et les métadonnées (van den Oord *et al.*, 2014 ; Zhang *et al.*, 2017). Le filtrage du pourriel peut également être considéré comme une forme de recommandation (ou de non-recommandation). Les techniques d'IA actuelles filtrent plus de 99,9 % du pourriel, et les services de courriel peuvent également recommander des destinataires potentiels, ainsi que préfigurer la réponse.

Jeu. Lorsque Deep Blue a vaincu le champion du monde d'échecs Garry Kasparov en 1997, les tenants de la supériorité de l'humain sur la machine ont placé leurs espoirs dans le jeu de go. Piet Hut, astrophysicien passionné du jeu de go, a pronostiqué qu'il faudrait « une centaine d'années avant qu'un ordinateur ne batte les humains au go – peut-être même plus longtemps ». Mais, à peine 20 ans plus tard, AlphaGo a surpassé tous les joueurs humains (Silver *et al.*, 2017). Ke Jie, le champion du monde, a déclaré : « L'année dernière, il était encore assez humain quand il a joué, mais cette année, il est devenu un dieu du go. » AlphaGo a bénéficié de l'étude de centaines de milliers de parties antérieures entre joueurs de go humains, ainsi que de l'expérience des experts de go qui étaient membres de l'équipe de conception.

Un programme ultérieur, AlphaZero, n'a utilisé aucune donnée humaine (sauf pour les règles du jeu), et a pu apprendre, en jouant uniquement contre lui-même, à vaincre tous ses adversaires, humains et machines, au go, aux échecs et au shogi (Silver *et al.*, 2018). Entre-temps, des champions humains ont été battus par l'IA à des jeux aussi divers que Jeopardy! (Ferrucci *et al.*, 2010), le poker (Bowling *et al.*, 2015 ; Moravcík *et al.*, 2017 ; Brown et Sandholm, 2019), ainsi que les jeux vidéo Dota 2 (Fernandez et Mahlmann, 2018), StarCraft II (Vinyals *et al.*, 2019) et Quake III (Jaderberg *et al.*, 2019).

Interprétation visuelle. Non contents de dépasser la précision humaine sur la tâche difficile ImageNet de reconnaissance d'objets, les chercheurs en vision par ordinateur se sont attaqués au problème encore plus difficile du légendage d'image, dont voici quelques exemples impressionnants : « une personne conduisant une moto sur un chemin de terre » ; « deux pizzas posées sur un fourneau » ; « un groupe de jeunes jouant à un jeu de Frisbee » (Vinyals, Toshev *et al.*, 2017). Les systèmes actuels sont cependant loin d'être parfaits : par exemple, un panneau d'interdiction de stationner partiellement masqué par de nombreux petits autocollants a été identifié comme « un réfrigérateur rempli de nourriture et de boissons ».

Médecine. Les algorithmes d'IA font maintenant aussi bien ou mieux que les médecins spécialistes pour le diagnostic de nombreuses pathologies, en particulier lorsqu'il est basé sur des images. En voici quelques exemples : maladie d'Alzheimer (Ding *et al.*, 2018), métastases cancéreuses (Liu *et al.*, 2017 ; Esteva *et al.*, 2017), troubles ophtalmiques (Gulshan *et al.*, 2016), et maladies de la peau (Liu, Jain *et al.*, 2019). Une étude systématique et une méta-analyse (Liu *et al.*, 2019) ont montré que les performances des programmes d'IA, en moyenne, étaient équivalentes à celles des professionnels de santé. L'une des priorités actuelles de l'IA médicale est de faciliter les partenariats homme-machine. Par exemple, le système Lyna atteint une précision globale de 99,6 % dans le diagnostic du cancer du sein métastatique (mieux qu'un expert humain non assisté), mais la combinaison homme-machine fait encore mieux (Liu *et al.*, 2018 ; Steiner *et al.*, 2018).

L'adoption généralisée de ces techniques est désormais limitée non pas par la précision du diagnostic, mais par la nécessité d'apporter la preuve de l'amélioration des résultats cliniques et de garantir la transparence, l'absence de biais et la confidentialité des données (Topol, 2019). En 2017, seules deux demandes d'IA médicale ont été approuvées par la FDA (Food and Drug Administration), mais ce chiffre est passé à 12 en 2018, et continue d'augmenter.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

Climatologie. Une équipe de scientifiques a remporté le prix Gordon Bell 2018 pour un modèle d'apprentissage profond qui permet de découvrir des informations détaillées sur des événements météorologiques extrêmes, informations qui étaient auparavant enfouies dans les données climatiques. Ils ont utilisé un supercalculateur avec du matériel spécialisé à base de GPU pour dépasser le niveau « exaop » (10^{18} opérations par seconde). C'est le premier programme d'apprentissage automatique à l'avoir fait (Kurth *et al.*, 2018). Rolnick *et al.* (2019) présentent un rapport de 60 pages répertoriant les différentes façons dont on peut utiliser l'apprentissage automatique pour lutter contre le changement climatique.

Ce ne sont là que quelques exemples des systèmes d'intelligence artificielle qui existent aujourd'hui. Ce n'est ni de la magie ni de la science-fiction, mais bien de la science, de l'ingénierie et des mathématiques, auxquelles ce livre fournit une introduction.

1.5 Risques et bénéfices de l'IA

Francis Bacon, le philosophe à qui l'on attribue la création de la méthode scientifique, a fait remarquer dans *The Wisdom of the Ancients* (1609) que « les arts mécaniques ayant leurs inconvénients, ainsi que leurs avantages, [ils] sont comme autant d'épées à deux tranchants qui servent tantôt à faire le mal, tantôt à y remédier ». Comme l'IA joue un rôle de plus en plus important dans tous les domaines : économique, social, scientifique, médical, financier et militaire, nous ferions bien de considérer les maux et les remèdes – en langage moderne, les risques et les bénéfices qu'elle peut apporter. Les sujets abordés ici sont traités en détail dans les chapitres 27 et 28.

Commençons par les bénéfices : pour dire les choses simplement, notre civilisation entière est le produit de notre intelligence humaine. Si on a accès à une intelligence mécanique sensiblement supérieure, le plafond de nos ambitions se trouve considérablement relevé. Le potentiel de l'intelligence artificielle et de la robotique pour libérer l'humanité du travail répétitif et subalterne et pour augmenter considérablement la production de biens et de services pourrait présager une ère de paix et d'abondance. La capacité à accélérer la recherche scientifique pourrait permettre de trouver des remèdes aux maladies et des solutions au changement climatique et à la pénurie de ressources. Comme Demis Hassabis, PDG de Google DeepMind, le suggère : « D'abord résoudre l'IA, ensuite utiliser l'IA pour résoudre tout le reste. »

Cependant, bien avant d'avoir la possibilité de « résoudre l'IA », on court le risque d'une utilisation impropre de celle-ci, que ce soit par inadvertance ou pour une autre raison. Certains de ces risques sont déjà apparents, tandis que d'autres semblent probables compte tenu des tendances actuelles :

- **Armes létales autonomes.** Les Nations unies les définissent comme des armes qui peuvent localiser, sélectionner et éliminer des cibles humaines sans intervention humaine. L'une des principales préoccupations concernant ces armes est qu'elles passent à l'échelle : comme elles ne nécessitent pas de supervision humaine, un petit groupe peut déployer un nombre arbitrairement élevé d'armes contre des cibles humaines, définies par n'importe quel critère de reconnaissance utilisable. Les technologies requises pour les armes autonomes sont analogues à celles des voitures autonomes. Des discussions informelles d'experts sur les risques potentiels des armes létales autonomes ont été initiées aux Nations unies en 2014 ; elles ont été poursuivies officiellement *via* la rédaction d'un traité préliminaire par un groupe d'experts gouvernementaux en 2017.
- **Surveillance et persuasion.** Bien que l'IA soit coûteuse, fastidieuse, et parfois juridiquement contestable du point de vue des experts en sécurité quand elle surveille les lignes téléphoniques, les flux de caméras vidéo, les courriels et autres canaux de messagerie, l'IA (reconnaissance de la parole, vision par ordinateur et compréhension du langage naturel) peut être utilisée pour surveiller massivement des individus et détecter des activités qui présentent un intérêt. Le comportement politique peut être modifié et contrôlé dans une certaine mesure en adaptant les flux d'information délivrés aux personnes par le biais des réseaux sociaux à l'aide de techniques d'apprentissage automatique – une préoccupation qui est devenue prégnante dès 2016 dans le cas des élections.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

- **Prise de décision biaisée.** La mauvaise utilisation, négligente ou délibérée, des algorithmes d'apprentissage automatique pour des tâches telles que l'évaluation des demandes de libération conditionnelle ou des demandes de prêt bancaire peut entraîner des décisions biaisées par la « race »[20], le genre ou d'autres critères de discrimination. Souvent, ce sont les données elles-mêmes qui reflètent des biais omniprésents dans la société.
- **Impact sur l'emploi.** Les inquiétudes concernant la suppression d'emplois par les machines sont vieilles de plusieurs siècles. L'histoire n'est jamais simple : les machines accomplissent certaines des tâches que les humains pourraient réaliser eux-mêmes, mais elles rendent également les humains plus productifs et donc plus employables, et améliorent la rentabilité des entreprises, qui sont alors capables de verser des salaires plus élevés. Elles peuvent aider certaines activités à être viables, alors qu'elles ne le seraient pas autrement. Leur utilisation entraîne généralement une augmentation de la richesse, mais a tendance à avoir pour effet de déplacer la richesse du travail vers le capital, ce qui aggrave encore les inégalités. Les progrès technologiques précédents, comme l'invention des métiers à tisser mécaniques, ont entraîné de graves perturbations sur le marché de l'emploi, mais les gens ont fini par trouver de nouvelles professions, même s'il est possible que l'IA affecte ces nouveaux métiers. Ce sujet est rapidement en train de devenir une préoccupation majeure pour les économistes et les gouvernements du monde entier.
- **Applications critiques en termes de sécurité.** À mesure qu'elles progressent, les techniques d'IA sont de plus en plus utilisées dans des applications à enjeux élevés et critiques pour la sécurité des personnes, comme la conduite automobile ou la gestion de l'approvisionnement en eau des villes. Des accidents mortels se sont déjà produits et mettent en évidence la difficulté de la vérification formelle et de l'analyse statistique des risques pour les systèmes développés à l'aide de techniques d'apprentissage automatique. Le domaine de l'IA devra développer des normes techniques et éthiques au moins comparables à celles qui prévalent dans d'autres disciplines telles que l'ingénierie ou les services de santé quand la vie des personnes est en jeu.
- **Cybersécurité.** Les techniques d'IA sont utiles pour se défendre contre les cyberattaques, par exemple en détectant des modèles de comportement inhabituels, mais elles contribuent également à augmenter la virulence, la survie et la capacité de prolifération des logiciels malveillants. Par exemple, des méthodes d'apprentissage par renforcement ont été utilisées pour créer des outils très efficaces pour les attaques automatisées et personnalisées par chantage et par hameçonnage.

Nous reviendrons plus en détail sur ces sujets en section 27.3. Les systèmes d'IA devenant de plus en plus performants, ils vont prendre une part croissante dans des rôles sociétaux tenus auparavant par des êtres humains. Puisque ces rôles ont été utilisés dans le passé pour commettre des méfaits, on peut s'attendre à ce que certains humains se servent des systèmes d'IA aux mêmes fins, avec plus d'efficacité. Tous les exemples donnés ci-dessus soulignent l'importance de la gouvernance et, à terme, de la réglementation. Actuellement, la communauté des chercheurs et les grandes entreprises impliquées dans la recherche en IA ont développé des principes d'autogestion volontaire pour les activités liées à l'IA (voir section 27.3). Les gouvernements et les organisations internationales mettent en place des organes consultatifs chargés de concevoir des réglementations appropriées pour chaque cas d'utilisation spécifique, de se préparer aux impacts économiques et sociaux et de tirer parti des capacités de l'IA pour résoudre les grands problèmes de la société.

Qu'en est-il à plus long terme ? Atteindrons-nous l'objectif visé depuis longtemps : la création d'une intelligence comparable ou supérieure à l'intelligence humaine ? Et, si nous y parvenons, que se passera-t-il ensuite ?

Pendant la majeure partie de l'histoire de l'IA, ces questions ont été éclipsées par le travail quotidien consistant à faire en sorte que les systèmes d'IA fassent quelque chose qui ressemble, même de loin, à de l'intelligence. Comme pour toute discipline de large ampleur, la grande majorité des chercheurs en IA se sont spécialisés dans un sous-domaine spécifique tel que le jeu, la représentation des connaissances, la vision ou la compréhension du langage naturel – souvent en partant du principe que les progrès dans ces sous-domaines contribueraient aux objectifs plus larges de l'IA. Nils Nilsson (1995), l'un des premiers responsables du projet Shakey au Stanford Research Institute (SRI), a rappelé à la communauté IA ces objectifs plus larges et a alerté sur le fait que les sous-domaines risquaient de devenir des fins en soi. Plus tard, certains fondateurs éminents de l'IA, parmi

20. NdT. Nous reprenons dans tout l'ouvrage le terme de « race » utilisé dans la version américaine, malgré son inadéquation, en raison de son utilisation statistique aux États-Unis.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

lesquels John McCarthy (2007), Marvin Minsky (2007) et Patrick Winston (Beal et Winston, 2009), se sont ralliés aux avertissements de Nilsson, en suggérant qu'au lieu de se concentrer sur des performances mesurables dans des applications spécifiques, l'IA devrait revenir à ses racines en s'efforçant, selon les termes de Herb Simon, de « créer des machines qui pensent, qui apprennent et qui créent ». Ils ont appelé cet objectif **IA de niveau humain (*human level AI* ou HLAI)**, correspondant à une machine qui devrait être capable d'apprendre à faire tout ce qu'un humain peut faire. Le premier symposium de ce courant a eu lieu en 2004 (Minsky *et al.*, 2004). Une autre initiative ayant des objectifs similaires, le mouvement **intelligence artificielle générale (*artificial general intelligence* ou AGI)**, a tenu sa première conférence en 2008 et publie depuis la revue *Journal of Artificial General Intelligence*.

À peu près au même moment, des voix s'élèvent contre la création d'une **superintelligence artificielle (*artificial superintelligence* ou ASI)**, une intelligence qui dépasserait de loin la capacité humaine, laissant entendre que ce pourrait être une mauvaise idée (Yudkowsky, 2008 ; Omohundro, 2008). Turing (1996) lui-même a fait le même constat lors d'une conférence donnée à Manchester en 1951, en puisant dans des idées antérieures de Samuel Butler (1863)[21] :

> Il semble probable qu'une fois que la méthode qui permettra aux machines de penser sera lancée, il ne lui faudra pas longtemps pour surpasser nos faibles capacités. […] À un moment donné, on devra donc s'attendre à ce que les machines prennent le contrôle, comme le mentionne Samuel Butler dans son livre *Erewhon*.

Ces préoccupations n'ont fait que s'amplifier avec les progrès récents dans le domaine de l'apprentissage profond, la publication de livres tels que *Superintelligence* par Nick Bostrom (2014), et les déclarations publiques de Stephen Hawking, Bill Gates, Martin Rees et Elon Musk.

Il est tout à fait naturel d'éprouver un sentiment général de malaise à l'idée de créer des machines superintelligentes. On pourrait appeler cela le **problème du gorille** : il y a environ sept millions d'années, un primate aujourd'hui éteint a évolué, avec une branche menant aux gorilles et une autre aux humains. Aujourd'hui, les gorilles ne sont pas très contents de la branche humaine : ils n'ont pratiquement aucun contrôle sur leur avenir. Si le succès de la création d'une IA surhumaine signifie que les humains cèdent le contrôle de leur avenir, alors peut-être devrions-nous arrêter les travaux sur l'IA et, en corollaire, abandonner les avantages qu'elle pourrait apporter. C'est l'essence même de l'avertissement de Turing : il n'est pas évident que nous puissions contrôler des machines plus intelligentes que nous.

Si l'IA surhumaine était une boîte noire venue de l'espace, il serait alors sage d'ouvrir la boîte avec prudence. Mais ce n'est pas le cas : *nous* concevons les systèmes d'IA ; par conséquent, s'ils finissent par « prendre le contrôle », comme le suggère Turing, ce sera le résultat d'un échec de conception.

Pour éviter un tel scénario, nous devons comprendre la source de l'échec potentiel. Norbert Wiener (1960), incité à envisager l'avenir à long terme de l'IA après avoir vu le programme de jeu de dames d'Arthur Samuel apprendre à battre son créateur, a déclaré :

> Si, pour réaliser nos objectifs, nous utilisons une machine dont le fonctionnement nous échappe, il vaut mieux être sûr que l'objectif placé dans la machine est celui que nous souhaitons vraiment.

De nombreuses cultures possèdent des mythes sur les humains faisant appel à des dieux, des génies, des magiciens ou des diables. Invariablement, dans ces histoires, ils obtiennent ce qu'ils demandent littéralement, puis le regrettent. Le troisième souhait, s'il y en a un, est de défaire les deux premiers. Nous appellerons cela le **problème du roi Midas**, par allusion au légendaire roi Midas de la mythologie grecque, qui, après avoir demandé que tout ce qu'il touche se transforme en or, l'a ensuite amèrement regretté après avoir touché sa nourriture, sa boisson et même les membres de sa famille[22].

Nous avons abordé cette question en section 1.1.5, où nous avons souligné la nécessité d'une modification significative du modèle standard par la mise en place d'objectifs fixes dans la machine. La solution au problème

21. Encore plus tôt, en 1847, Richard Thornton, rédacteur en chef de la revue *Primitive Expounder*, s'était élevé contre les calculatrices mécaniques : « L'esprit […] se dépasse et se débarrasse de la nécessité de sa propre existence en inventant des machines pour penser à sa place. […] Mais qui sait si ces machines, une fois perfectionnées, ne pourraient pas trouver un plan pour remédier à tous leurs défauts et ensuite sécréter des idées hors de portée de l'esprit des simples mortels ! »

22. Midas aurait mieux fait de suivre les principes de base de la sécurité et d'inclure un bouton « *undo* » et un bouton « pause » dans son souhait.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

de Wiener est de ne pas avoir d'« objectif fixe placé dans la machine » du tout. Ce que nous voulons, ce sont des machines qui s'efforcent d'atteindre des objectifs humains, mais qui savent qu'elles ne savent pas exactement quels sont ces objectifs.

Il est peut-être regrettable que presque toutes les recherches sur l'IA menées jusqu'à présent l'aient été dans le cadre du modèle standard, ce qui signifie que presque tout le matériel technique de cette édition est le reflet de ce cadre intellectuel. On commence cependant à avoir des premiers résultats correspondant au nouveau mode de pensée. Au chapitre 16, nous montrons qu'une machine est incitée à accepter de se désactiver si et seulement si elle est incertaine de l'objectif humain. Au chapitre 18, nous formulons et étudions les **jeux d'assistance** (***assistance games***), qui décrivent mathématiquement des situations où une personne possède un objectif et où une machine essaie de l'atteindre, mais est initialement incertaine à propos de cet objectif. Dans le chapitre 22, nous expliquons les méthodes d'**apprentissage par renforcement inverse** (***inverse reinforcement learning***), qui permettent aux machines d'en savoir plus sur les préférences humaines à partir de l'observation des choix que font les humains. Au chapitre 27, nous explorons deux des principales difficultés rencontrées : premièrement, nos choix dépendent de nos préférences par l'intermédiaire d'une architecture cognitive qui est difficile à inverser ; deuxièmement, les êtres humains n'ont peut-être pas de préférences cohérentes en premier lieu, que ce soit individuellement ou en tant que groupe – il n'est donc pas nécessairement évident de savoir ce que les systèmes d'IA doivent faire pour nous.

Résumé

Ce chapitre a défini l'IA et établi le contexte culturel et historique de son développement :

- Chaque personne aborde l'IA avec un objectif différent en tête. Les deux questions essentielles qui se posent sont les suivantes : vous intéressez-vous plutôt à la pensée ou au comportement ? Voulez-vous modéliser le comportement humain ou essayer d'obtenir des résultats optimaux ?
- Selon ce que nous avons appelé le modèle standard, l'IA a principalement trait à l'**action rationnelle**. Un **agent intelligent** idéal entreprend la meilleure action possible dans une situation donnée. C'est dans cette perspective que nous étudions le problème de la construction d'agents intelligents.
- Deux améliorations de cette proposition élémentaire sont nécessaires : premièrement, la capacité de tout agent, humain ou autre, à choisir des actions rationnelles est limitée par l'impraticabilité du calcul pour ce faire ; deuxièmement, le concept de machine qui poursuit un objectif précis doit être remplacé par celui de machine qui poursuit des objectifs au bénéfice des humains, en restant dans l'incertitude quant à la nature de ces objectifs.
- Les philosophes (dès l'an 400 av. J.-C.) ont rendu l'IA concevable en suggérant que l'esprit ressemble à certains égards à une machine, qu'il opère sur des connaissances encodées dans un langage interne et que la pensée peut faciliter le choix des actions à entreprendre.
- Les mathématiciens ont fourni les outils permettant de manipuler les énoncés logiques de certitude ainsi que les énoncés probabilistes d'incertitude. Ils ont également jeté les bases de la compréhension du calcul et du raisonnement sur les algorithmes.
- Les économistes ont formalisé le problème de la prise de décision qui maximise l'utilité espérée pour le décideur.
- Les neurobiologistes ont fait certaines découvertes sur le fonctionnement du cerveau, et sur ses similitudes et ses différences avec un ordinateur.
- Les psychologues ont adopté l'idée que les humains et les animaux peuvent être vus comme des machines de traitement de l'information. Les linguistes ont montré que l'usage du langage s'insère dans ce modèle.
- Les informaticiens ont développé des machines de plus en plus puissantes qui rendent possibles les applications de l'IA, et les ingénieurs logiciels ont fait en sorte qu'elles soient plus utilisables.
- La théorie du contrôle traite de la conception de dispositifs opérant de manière optimale à partir du *feedback* fourni par l'environnement. À l'origine, les outils mathématiques utilisés par cette discipline étaient différents de ceux de l'IA, mais ces deux domaines sont en train de se rapprocher l'un de l'autre.

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet

- L'histoire de l'IA se caractérise par des cycles de succès, d'optimisme déplacé, puis de perte d'enthousiasme et de financement (qui résultent des étapes précédentes). Il y a également eu des cycles d'introduction de nouvelles approches créatives et de perfectionnement systématique des meilleures approches connues.
- L'IA a considérablement mûri par rapport à ses premières décennies, tant sur le plan théorique que méthodologique. À mesure que les problèmes traités par l'IA sont devenus plus complexes, ce domaine est passé de la logique booléenne au raisonnement probabiliste, et de la connaissance artisanale à l'apprentissage automatique à partir de données. Cela a conduit à l'amélioration des capacités des systèmes réels et à une plus grande intégration à d'autres disciplines.
- À mesure que les systèmes d'IA trouvent leur application dans le monde réel, il est devenu nécessaire de se préoccuper d'un large éventail de risques et de conséquences éthiques.
- À plus long terme, nous sommes confrontés au difficile problème du contrôle des systèmes d'IA superintelligents, qui peuvent évoluer de manière imprévisible. La résolution de ce problème semble nécessiter un changement dans notre conception de l'IA.

Notes bibliographiques et historiques

Nils Nillson, l'un des pionniers du domaine de l'IA, en propose un historique complet (Nilsson, 2009). Pedro Domingos (2015) et Melanie Mitchell (2019) offrent un aperçu de l'apprentissage automatique destiné au grand public. Kai-Fu Lee (2018) décrit la course au leadership international en matière d'IA. Martin Ford (2018) interroge 23 éminents chercheurs en IA.

Les principales sociétés savantes dans le domaine de l'intelligence artificielle sont l'association AAAI (Association for the Advancement of Artificial Intelligence), le groupe SIGAI (ACM Special Interest Group in Artificial Intelligence), anciennement SIGART, la European Association for AI et la société AISB (Society for Artificial Intelligence and Simulation of Behaviour). Le Partenariat mondial sur l'IA réunit de nombreuses organisations aussi bien commerciales qu'à but non lucratif concernées par l'éthique et les impacts sociaux de l'IA. Le magazine de l'AAAI contient de nombreux articles thématiques et didactiques, et son site web, `aaai.org`, présente des actualités, des tutoriels et des informations générales.

Les travaux les plus récents sont publiés dans les comptes-rendus des principaux colloques consacrés à l'IA : la conférences IJCAI (*International Joint Conference on AI*), la conférence annuelle ECAIl (*European Conference on AI*) et la conférence de l'AAAI. Les conférences *International Conference on Machine Learning* et *Neural Information Processing Systems* couvrent l'apprentissage automatique. Les principales publications généralistes en IA sont *Artificial Intelligence*, *Computational Intelligence*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Intelligent Systems* et la publication en ligne *Journal of Artificial Intelligence Research*. On trouve aussi de nombreuses conférences et publications consacrées à des domaines spécifiques (voir les chapitres concernés).

© 2021 Pearson France - Intelligence artificielle, 4e édition
Stuart Russell, Peter Norvig - Traduit par Laurent Miclet, Fabrice Popineau, Claire Cadet