

Introduction

Le présent ouvrage aborde la thématique d'une nécessaire éthique accompagnant les développements de l'intelligence artificielle (IA), quels que soient les domaines concernés. En effet, l'intelligence artificielle est en train de modifier en profondeur nos pratiques, mais aussi la manière dont nous pensons ou interagissons, tant à l'échelle des individus et des organisations que des sociétés, tant dans notre vie privée que publique ou professionnelle.

Toutefois, en guise de préambule, il nous paraît important d'essayer de cerner les contours de ce secteur d'activité de plus en plus présent dans nos usages quotidiens comme professionnels. À ce niveau, nous pouvons observer l'existence de nombreuses définitions de l'IA qui rendent difficile la démarche consistant à en circonscrire précisément les contours.

Cette multiplicité d'approches de l'IA peut originellement s'expliquer par sa genèse pluridisciplinaire. On considère en effet que l'acte de naissance officiel de l'IA correspond aux conférences Macy à New York entre 1942 et 1953, regroupant mathématiciens, logiciens, anthropologues, psychologues et économistes ayant pour objectif d'étudier de quelles manières pourrait-on reproduire, par l'informatique, l'apprentissage et l'intelligence humaine.

Il faut aussi, dans ce contexte de réflexion sur la définition de l'IA, tenir compte des deux principaux axes structurels de cette notion, à partir des années 1950, à savoir :

- un axe centré sur une approche symbolique qui est prégnante jusque dans les années 1980, et qui aborde la notion d'intelligence par l'utilisation et la

manipulation de symboles à partir de règles formelles (« systèmes symboliques physiques » de Newell et Simon en 1976) ;

– un axe dit « connexionniste » qui s'inspire des réseaux neuronaux biologiques et dont le fondement repose sur le fait qu'une perception ou un comportement intelligent est la conséquence d'interactions entre diverses unités computationnelles gérées par des règles d'apprentissage (voir les travaux pionniers de McCulloch et Pitts dès 1943, ou plus proche de nous, ceux de Levesque en 2013 qui définit l'IA comme une approche abordant les comportements intelligents dans une logique computationnelle).

Cette notion fait débat parmi des spécialistes du domaine (Lecun 2019), elle est même régulièrement remise en question, en mettant en exergue plutôt les capacités de reconnaissance par l'apprentissage de la machine ou du dispositif étudié, que par l'intelligence elle-même.

Certains spécialistes reconnus dans ce domaine considèrent même que l'IA n'est pas si intelligente qu'on le dit et que ce n'est pas l'intelligence qui caractérise les systèmes d'IA d'aujourd'hui, mais leur capacité de reconnaissance grâce à l'apprentissage machine. À partir de cette approche, certains auteurs préfèrent parler d'intelligence augmentée (Carter et Nielsen 2017) ou vont même jusqu'à dire que l'IA n'existe pas, dans la mesure où le concept d'intelligence n'est pas défini précisément et que le débat reste très ouvert à ce niveau (Julia 2019 ; Karppi et Granat 2019).

La littérature propose trois familles de définitions qui sont autant de facettes complémentaires de ce qu'est l'IA et de ce que l'IA peut réaliser. La première famille s'appuie sur une compréhension technique ou technologique de l'IA et la présente comme une branche des sciences du numérique (ou de l'informatique).

Pour certains, l'IA se définit comme une famille de modèles statistiques, à savoir des modèles de prédiction à partir du traitement de données, utilisant des techniques de *machine learning*. Selon cette définition, l'IA est une technique, car elle décrit un ensemble de procédés (algorithmes) ou de méthodes susceptibles d'obtenir un résultat déterminé. L'IA est aussi une technologie en ce qu'elle est une discipline d'étude des techniques spécifiques à l'IA et donc peut se comprendre comme un domaine qui permet de développer de nouvelles techniques.

La limite de définitions d'inspiration exclusivement technique ou technologique consiste à restreindre le champ des possibles et à se désintéresser d'évolutions fonctionnelles susceptibles de faire émerger de nouvelles techniques. Elle nous oblige à regarder l'IA par le petit bout de la lorgnette.

Une seconde famille définit l'IA comme une machine capable de reproduire ou simuler l'intelligence humaine (c'est-à-dire apprendre, comprendre, raisonner, interagir, etc.). Ainsi, une science de l'IA serait l'étude des moyens permettant d'automatiser les processus de perception, d'apprentissage, de raisonnement, de décision et d'action.

Le Bureau international des brevets (WIPO) va dans ce sens lorsqu'il définit l'IA comme « une branche de l'informatique qui a pour objet de concevoir des machines et des systèmes à même d'accomplir des tâches faisant appel à l'intelligence humaine, avec une intervention humaine limitée ou nulle ».

Toutefois, souligner la capacité d'une intelligence artificielle à simuler ou reproduire des processus de délibération, de décision et d'action humaines permet à un public large de s'approprier le concept et de laisser libre cours à l'imagination de futurs possibles. Ces futurs imaginés peuvent aller des fantasmes les plus extravagants ou de peurs injustifiées, aux anticipations potentiellement productives et aux prises de conscience nécessaires à la définition de règles d'encadrement.

Une troisième famille de définitions adopte une perspective plus fonctionnelle. Elle présente l'IA comme un ensemble de systèmes artificiels capables de s'adapter de façon plus ou moins autonome à son environnement, pour parvenir à des fins prédéfinies. La définition proposée par l'OCDE dans son rapport sur « L'intelligence artificielle dans la société (2019) » va dans ce sens : « un système d'intelligence artificielle (ou système d'IA) est un système automatisé qui, pour un ensemble donné d'objectifs définis par l'homme, est en mesure d'établir des prévisions, de formuler des recommandations, ou de prendre des décisions influant sur des environnements réels ou virtuels. »

Cette définition s'inspire de l'approche initiale proposée par Simon en 1969 dans son ouvrage *Sciences of the Artificial* et met en avant la capacité d'interaction avec un environnement sur lequel le système d'IA peut agir et donc contribuer à modifier. Cette définition permet aussi d'envisager différentes

classes de systèmes selon la « fermeture » ou « l'ouverture » d'une fin prédéfinie (l'objectif assigné à la machine). De fait, il est alors possible d'envisager et d'étudier différentes classes de comportements, sans se contraindre à des considérations techniques particulières.

Cette conception fonctionnelle – ou fonctionnaliste – met donc l'accent sur les caractéristiques comportementales. Elle permet d'aborder l'étude des systèmes IA non seulement comme des systèmes distincts, identifiables, et dont la fonction est prédéfinie, mais aussi comme un ensemble de sous-systèmes en interaction et dont le comportement global n'est pas toujours prévisible.

Cette question devient particulièrement pertinente avec la multiplication des formes de systèmes, l'explosion des données produites (notamment du fait de la multiplication des objets connectés) et immédiatement disponibles, ainsi que la progressive délocalisation des algorithmes IA au plus près des objets producteurs de données (*via* le Edge Computing).

Cette dernière conception pourrait se rapprocher de la méthode qui a été privilégiée par la Commission européenne et qui repose sur une approche par les risques (21 avril 2021). En effet, cette récente proposition de cadre réglementaire est fondée sur une définition de l'IA à l'épreuve du temps, selon une approche fondée sur les risques (inacceptables, élevés, limités, minimes). L'article 3 du projet de règlement définit le système d'intelligence artificielle comme un logiciel qui peut, pour un ensemble donné d'objectifs définis par l'homme, générer des résultats tels que du contenu, des prévisions, des recommandations, ou des décisions influençant les environnements avec lesquels ils interagissent.

Ces trois familles de définitions se complètent et peuvent être combinées pour appréhender les questions relatives aux impacts de l'IA sur la société, l'économie et notamment l'éthique. Il est important aussi d'envisager la structuration des études scientifiques autour de l'IA. Par exemple, la possibilité de ces interactions entre systèmes hétérogènes nous ramène à des problématiques proches de l'étude des comportements de systèmes biologiques, des animaux ou des sociétés animales par exemple. C'est ainsi qu'un ensemble de scientifiques a initié des travaux utilisant des techniques expérimentales s'inspirant de l'éthologie pour analyser l'IA (voir l'article « Machine Behavior » paru dans *Nature* en 2019).

Dans ce contexte, et en intégrant les performances atteintes par les machines apprenantes actuelles, notre approche et vision de l'IA doit prendre en compte l'éventuelle survenue d'une IA « consciente » donc équivalente dans un premier temps à celle de l'humain (Goertzel 2014) et étant à même de la dépasser (Boström 2014). Nos définitions de l'IA ne peuvent évacuer les réflexions sur une l'IA dite « forte » par opposition à une IA « faible » correspondant aux développements actuels de cette notion¹. Cette réflexion permet d'aborder l'IA sous un angle différent de celui basé sur la reproduction de l'intelligence humaine (ou d'autres formes d'intelligence présentes chez certaines espèces animales). Dans le contexte d'une IA dite « forte », une machine ou un dispositif dont l'intelligence serait égale, voire supérieure à celle d'un être humain, pourrait alors, afin de réaliser un objectif, utiliser tous les moyens à sa disposition sans tenir compte de notions éthiques et de ce que Russell (2019) appelle la « structure de préférence ». Cette dernière ne prend en compte que les solutions compatibles avec l'être humain et qui pour lui est implicite (dans le cas bien sûr où les intentions sont bonnes).

L'IA représente donc l'un des enjeux sociétaux majeurs dont les implications et les applications occupent une place de plus en plus prégnante dans nos modes de vie, et dans le cadre de développements technologiques liés à la collecte et à l'analyse de données massives. Par les outils qu'elle développe, l'IA peut améliorer considérablement notre futur, mais dans le même temps entraîner le développement de nombreux risques et défis comportant potentiellement d'inquiétantes zones d'ombres, notamment au niveau de la vie démocratique et l'usage de données personnelles éminemment sensibles.

Après s'être développée pendant des années, en ignorant complètement ces dimensions, et en faisant de l'IA une vaste expérimentation sociale, éthique et politique sur laquelle les citoyens ainsi que les associations et la grande majorité

1. Certains auteurs comme Collins (2018) proposent une typologie plus fine des IA. Ce dernier distingue six niveaux d'IA : le premier fait référence aux technologies actuelles, le deuxième correspond à la notion de « prothèses asymétriques » qui peuvent réaliser une activité humaine (Alexa de Google ou encore les chatbots) sans pour autant être capables de repérer et pallier nos erreurs. Les quatre niveaux suivants sont de l'ordre d'une certaine vision prospective « prothèses symétriques » corrigeant les erreurs d'une personne ; machines et dispositifs capables de reproduire totalement ou partiellement les processus internes de l'intelligence humaine ; machines autonomes anthropomorphiques pouvant former des communautés voire des sociétés ; machines ayant une corporéité différente de nous, pouvant s'autoaméliorer au fil des générations et vivre également en société.

des entreprises n'avaient pas été consultées, il est devenu évident que les problèmes éthiques se trouvent aujourd'hui au cœur des enjeux cruciaux de l'IA : biais, discriminations, droits des personnes, n'en constituent que quelques exemples.

De fait, la conception et l'usage de ces nouvelles technologies doivent être accompagnés d'une réflexion éthique sérieuse. Il convient d'ailleurs de préciser que sur un plan international, de nombreux acteurs sociétaux (citoyens, ONG, institutions, entreprises, etc.) confirment cette urgente nécessité de réfléchir éthiquement à l'IA, pouvant amener à l'élaboration d'une éthique de (ou pour) l'IA. Cette réflexion suscite un certain nombre de questions comme le fait de se demander si l'IA relève de l'éthique générale ou si elle est spécifique. Si elle est spécifique, n'observe-t-on pas une forte prégnance d'une éthique conséquentialiste souvent trop utilitariste ? Ne doit-on pas élargir le raisonnement autour de l'éthique de l'IA en intégrant d'autres démarches éthiques à l'instar de celles présentes en philosophie morale (déontologisme, arétaïsme et axiologisme) ?

Un autre questionnement important qui se pose aujourd'hui est de penser les contours d'une éthique par design de l'IA, qui ne se limite, ni à appliquer de grands courants théoriques ou de grands principes qui auraient été pensés en amont, ni à évaluer, cette fois en aval, les impacts ou conséquences. Cette éthique par design située à chaque moment du processus de conception, développement, diffusion des IA et retour d'expériences, propose une dynamique réflexive entre théories éthiques et cas concrets ou pratiques, en permettant de rendre visibles ou audibles des points de vue ou positions non majoritaires, en sus de ceux communément adoptés et pouvant être par ailleurs contradictoires.

L'objectif principal de cet ouvrage divisé en trois parties consiste à tenter d'apporter un éclairage sur cette problématique de l'éthique de l'IA à l'aune de différents champs disciplinaires et secteurs d'activité.

La première partie a une dimension plus conceptuelle et renvoie à l'imaginaire de l'IA. À ce niveau, Nicolas Oliveri aborde l'IA par le biais des humanités robotiques. Il part du constat que les problématiques de recherche autour des humanités numériques se retrouvent et se développent autour de disciplines scientifiques diverses. Sur un plan épistémologique, ce concept est encore en mouvance et renvoie au questionnement du devenir humain face à

la technologisation du monde. Dans ce contexte, sur un plan exploratoire et interculturel, les « humanités robotiques » présentent un futur imaginaire, au sein duquel la technologie robotique revêt une dimension à majorité salvatrice (voir surtout le cinéma américain, et le manga ou l'animation japonaise). De fait, Nicolas Oliveri présente une déconstruction analytique des mécanismes de fabrication de cette vision utopique de la pensée post-humaniste dans laquelle de fortes représentations sociales ont un rôle essentiel dans l'étude de la fusion homme-machine au regard de l'IA, que ce soit par le biais des « réalisateurs-créateurs » ou des « spectateurs-récepteurs ».

Dans cette partie, Juan Carrasco réfléchit sur les frontières de la technologie et l'avenir de l'IA. Il part du constat de l'utilisation accrue de l'IOT (*Internet Of Things*) par les individus faisant appel à des environnements numériques complexes. Le degré de précision et la puissance de traitement de l'IA associés aux activités humaines entraînent une certaine remise en question et évolution des frontières traditionnelles de la technologie et de la communication. Cette mutation nous astreint à nous questionner sur les implications éthiques de ces usages, point d'orgue de son chapitre. L'auteur développe une approche basée sur trois concepts-clés pour mieux appréhender les limites de la technologie et les ramifications de l'intelligence artificielle. Le point central de sa réflexion part de la manière dont la politique de l'archive de Derrida croise l'exploration de Marcuse et de Heidegger, à propos des fondements de la technologie moderne, et ceci, afin de mieux saisir les implications éthiques que crée la forte évolution de l'IA générative dans le domaine de la communication. Juan Antonio Carrasco nous incite ainsi à réimaginer nos approches des interactions entre individus et de leur coexistence avec la multiplicité d'objets « intelligents » composant notre environnement, et ce, au-delà des discours, voire des invocations « technologistes » et « millénaristes », en l'ayant plus largement présent à l'esprit, et en considérant leurs incidences sur notre système démocratique et nos structures sociétales.

Pour conclure cette partie, Alexandre Carbonneau traite de l'impact des évolutions technoscientifiques sur l'écologie mentale. En effet, des entreprises à l'instar des GAFAM (ou GAMMA depuis que Facebook se nomme Meta) ou des BATX en Chine, mais également des États, investissent dans des expérimentations ouvrant la voie à un *homo silicium* aux compétences logico-mathématiques accrues par le biais d'implants cérébraux. Ses réflexions portent sur l'impact d'une telle neuro-augmentation, sur la perception que nous avons

de nous-mêmes, mais également, de l'autre et de nos sociétés. Dans ce chapitre, Alexandre Carbonneau concentre son attention sur les nouvelles relations et interactions entre l'individu et l'écosystème technique et technoscientifique qu'il a contribué à créer. Il précise que jusqu'à la fin du siècle dernier, les innovations imaginées et développées par les humains lui ont permis de s'extraire et dépasser sa condition naturelle, tout en gardant la maîtrise sur la rationalité de son système technique. Or, depuis cette période, notamment au cours de ce siècle, par l'incidence d'un rythme d'innovations technologiques (notamment celles centrées sur les nanotechnologies) toujours plus soutenu, ce constat tend à s'invalider. Cet état de fait contribue à créer une forme de sanctification irrationnelle des sciences et des techniques associées, pouvant renvoyer à un nouvel ordre sociétal centré sur le mythe (de plus en plus réaliste) de l'homme « augmenté » et de fait, pourvu d'une surpuissance intellectuelle et physique.

La seconde partie met en exergue la nécessité de la mise en place d'un cadre éthique dans le contexte d'usages intégrants. Les deux chapitres qui la composent illustrent cette « obligation éthique », par le biais de deux secteurs d'activité en plein essor.

Dans cette partie, André Boyer et Faranak Farzaneh traitent des incidences de l'IA sur l'amour, et plus spécifiquement de la place des sites de rencontres dans les relations amoureuses. Ces sites témoignent de la « plateformesation » croissante de nos vies, à l'instar du e-commerce. Les auteurs considèrent ce fort sentiment affectif comme amenant à des comportements spécifiques, et il peut être considéré comme un élément-clé dans les relations sociales, car il intègre non seulement les normes et les valeurs sociales, mais aussi et surtout les sentiments des personnes. De ce fait, comme pour la plupart de nos interactions sociales qui évoluent de plus en plus dans des espaces hyper connectés, nos relations amoureuses, d'après les auteurs de ce chapitre, sont à la fois soutenues, mais également questionnées par l'évolution des réseaux sociaux associés aux plateformes digitales qui nous incitent à réaliser un ensemble de choix similaires au comportement d'un consommateur, afin d'identifier la personne en phase avec nos attentes et désirs. Pour les auteurs, nous assistons à une remise en question du modèle « coutumier » de l'amour romantique en tant que cadre prégnant des rapports amoureux. Dans ce contexte, par les biais des rencontres en ligne, nous assistons à un effacement total de l'imprévu et du premier « face à face », ainsi que des émotions qui y sont associées. André

Boyer et Faranak Farzaneh analysent donc l'impact du numérique sur les relations amoureuses et leur perception, notamment par le biais des réseaux sociaux, des sites de rencontres et de fait, des IA associées à ces supports, ceci par le prisme d'une théorie de l'amour.

Le deuxième chapitre composant cette partie traite de l'importance de l'IA dans la conception et le développement des *smart cities*, et les réflexions éthiques qui y sont associées. Franck Debos rappelle que cette vision de la cité se fonde sur une interconnexion des réseaux en croisant les données émises par chacun d'entre eux afin de mettre en place un nouvel écosystème urbain devant répondre aux évolutions sociétales. L'auteur précise que les mutations actuelles des villes, qu'elles soient organisationnelles, technologiques ou sociétales, doivent intégrer les problématiques inhérentes à la transition écologique, mais aussi être associées à des réflexions sociales. Dans ce cadre, nous devons retrouver les fondements du développement durable au sein de l'approche globale initiée par la *smart city*, et ce, afin de développer une gouvernance participative, une gestion expérimentée des ressources naturelles, le tout dans une optique de satisfaction des besoins et attentes des institutions, des entreprises, des associations et des citoyens. L'auteur présente ici les enjeux et voies d'action majeurs pour faire évoluer la démarche fortement technosolutionniste qui domine dans les *smart cities*, vers une approche permettant à ces territoires urbains d'avoir un impact positif sur un plan environnemental et social. Dans ce contexte, la place de l'IA est de plus en plus prégnante et il est nécessaire de réfléchir à un cadre éthique des usages de l'IA au sein des projets de *smart cities*, afin de mieux appréhender les risques de dérives et de bien mettre la dimension humaine au centre de cette vision digitalisée de la cité.

Le dernier chapitre de cette partie s'inscrit dans la continuité du précédent en termes de réflexions et de terrain d'étude. Dans ce chapitre, Maria Daniella Lancini et Franck Debos partent du constat qu'au niveau des articles, ouvrages, livres blancs, etc., traitant des facteurs de réussite d'une *smart city*, il est assez peu traité de l'importance du rôle des entreprises à ce niveau. Il s'agira ici de déterminer les principaux enjeux et attentes des entreprises françaises vis-à-vis des *smart cities* et ce, indépendamment de leur secteur d'activité ou de leur taille. Les auteurs se penchent en parallèle sur l'évolution et la mise en perspective de ces mêmes entreprises lorsque celles-ci ont intégré l'écosystème d'une *smart city*. Ces différentes réflexions permettront de définir la valeur

de leur impact au niveau de la mutation d'une ville « traditionnelle » en une *smart city*. Dans ce contexte où l'IA est omniprésente, les auteurs traiteront également de la nécessité d'une IA réglementée afin de développer les relations entre les acteurs considérés.

La dernière partie de cet ouvrage convoque l'éthique de l'IA au niveau de communautés professionnelles qui subissent des mutations importantes par la digitalisation de leurs activités. À ce niveau, Loïc Sumien aborde l'impact de l'IA dans l'éducation aux médias et à l'information. En effet, l'IA qui existe depuis des décennies dans le secteur des médias et associée aux réseaux sociaux numériques, restructure de façon importante les modes de communication et d'information des individus, et plus largement l'écosystème de l'information par le biais de nouveaux supports de production et diffusion accessibles à tous. L'auteur précise par ailleurs que l'intérêt que suscite l'IA est très récent et peut être associé à l'émergence d'outils apparus depuis deux ans (ChatGPT, Midjourney, etc.) et ouvre une réflexion sur des usages et apprentissages éthiques de cette technologie. Afin d'illustrer les enjeux soulevés par l'IA, Loïc Sumien étudie le secteur de l'éducation aux médias et à l'information traversé par de nouvelles pratiques pédagogiques associées à l'IA, et qui se développent en relation forte avec les sciences de l'information et de la communication. Ces pratiques mettent en exergue la nécessité d'en acquérir un usage éthique, expérimenté et réfléchi, ainsi que l'importance pour l'IA de poursuivre sa réinvention et répondre aux attentes des publics concernés.

De façon concomitante, Horea Mihai Badau se penche sur les algorithmes d'IA, présents à notre insu dans nos smartphones, et qui agissent en tant que « médiateurs » actifs en termes de communication numérique, par la sélection et l'interprétation de l'information, à l'instar des éditeurs de journaux télévisés. L'auteur a créé un simulateur d'algorithmes d'IA similaire au système de recommandation de TikTok, et a développé une interface utilisateur afin de rendre les interactions plus aisées. Plusieurs principes de la « pensée algorithmique » émergent, associés aux usages des utilisateurs, à l'optimisation d'objectifs donnés et à la dépendance aux données historiques. Horea Mihai Badau a ensuite procédé à une analyse des nœuds (dans ce cas, les comptes d'influenceurs) en adéquation avec les algorithmes, ainsi qu'une analyse des incidences des algorithmes d'IA sur les univers sémantiques des utilisateurs, par le biais des motifs et thèmes récurrents existants dans leurs interactions. Cette recherche a permis d'identifier les motivations des individus les plus

impactées par les algorithmes et met en lumière la nécessité d'un cadre éthique de l'IA à ce niveau.

La dernière contribution de cette partie est présentée par Vanessa Landaverde Kastberg et Matina Magkou et traite des défis éthiques des usages de l'IA dans les Industries culturelles créatives (ICC), ainsi que les débats et réponses qu'ils suscitent au sein d'organismes internationaux, en l'occurrence l'UNESCO et l'UE, et ce, à partir des bases conceptuelles et des points de départ de ces institutions en termes d'engagements pour une IA éthique. Les autrices réfléchissent également sur les rapports entre ce domaine technologique, et les besoins et droits des créatifs au sein des ICC. Elles adossent leur démarche sur des recherches documentaires se rapportant au questionnement éthique à propos de l'IA, et au niveau des deux institutions précitées. Ce travail est enrichi et complété par des entretiens auprès de représentants de ces organismes ainsi que de groupes d'intérêts au niveau des ICC, mais également des réflexions collectives qui émanent de colloques internationaux traitant des rapports entre l'IA et les ICC.

Bibliographie

- Boström, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford.
- Carter, S., Nielsen, M. (2017). *Using Artificial Intelligence to Augment Human Intelligence*. Distill.
- Collins, H. (2018). *Artificial Intelligence. Against Humanity's Surrender to Computers*. Polity Press, Cambridge.
- Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, 5(1), 1–46.
- Julia, L. (2019). *L'intelligence artificielle n'existe pas*. First, Paris.
- Karppi, T., Granata, Y. (2019). Non-artificial non-intelligence: Amazon's Alexa and the frictions of AI. *AI & Society*, 34(4), 867–876.
- Lecun, Y. (2019). *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*. Odile Jacob, Paris.
- Levesque, H. (2013). On Our Best Behaviour. Dans *IJCAI*. Pékin, Chine.

- McCulloch, W.S., Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5, 115–133.
- Newell, A., Simon, H.A. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3), 113–126.
- Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.