Entre rêves et illusions… L'intelligence artificielle en question

Isabelle Linden

Dans Revue d'éthique et de théologie morale 2020/3 (N° 307), pages 11 à 27
Éditions Éditions du Cerf

ISSN 1266-0078
ISBN 9782204138086
DOI 10.3917/retm.310.0011

ÉTHIQUE ET INTELLIGENCE ARTIFICIELLE

Isabelle Linden

Entre rêves et illusions... L'intelligence artificielle en question

Introduction

L'intelligence artificielle (IA) est partout, elle réussit, elle échoue, elle fait rêver, elle effraie… Les frontières sont parfois floues entre mythe et réalité, entre science-fiction et futurologie. L'IA pose de multiples questions, non seulement dans le champ technique, mais plus encore quant à son intégration dans la vie quotidienne et le projet collectif. C'est une responsabilité collective que de penser ces questions pour aujourd'hui et d'anticiper celles que l'IA posera demain ; elle comporte des aspects techniques mais aussi des questions sociales, anthropologiques, éthiques, juridiques… et demande une approche largement transdisciplinaire. Les réflexions qui suivent ambitionnent de fournir quelques bases qui puissent permettre aux chercheurs des différentes disciplines d'intégrer les réflexions et débats que suscite l'intelligence artificielle.

Cet article commence par un exposé technique qui vise à introduire en des termes accessibles aux non-informaticiens les fondements des principales techniques d'intelligence artificielle. Les caractéristiques de ces techniques sont alors développées de façon à dégager les limites et les ambitions que peuvent avoir les techniques actuelles. Dans un deuxième temps, nous discutons jusqu'à quel point l'IA peut être qualifiée d'intelligente. Ce questionnement est prolongé par quelques éléments à propos de la responsabilité et de la moralité de l'IA. Enfin, nous interrogeons le projet de société auquel l'IA peut contribuer. En conclusion, une invitation est lancée à l'exploration transdisciplinaire des questions abordées.

Que recouvre l'expression « intelligence artificielle » ?

Pour cerner de façon pertinente les problématiques soulevées par l'intelligence artificielle, il importe de bien comprendre l'objet dont il est question. La dénomination « intelligence artificielle » a été adoptée lors du congrès de Dartmouth en 1956. Cette réunion scientifique ambitionnait d'étudier les techniques qui permettent à une machine de simuler différentes facultés de l'intelligence.

Différentes approches s'inscrivent dans cette tradition. La plupart d'entre elles conçoivent un cerveau artificiel comme un système fonctionnant suivant les règles d'un algorithme. Il reçoit des ensembles de signaux (en entrée) : lumières, couleurs, températures, positions, données issues d'un formulaire, d'une base de données ou des capteurs d'un robot, et propose (en sortie) une réponse : des informations ou des actions.

Les premiers grands succès de l'IA, dans les années 1970, concernaient des systèmes d'aide à la décision ou systèmes experts dans des domaines très variés tels que le support au diagnostic médical, la prédiction d'achat ou d'attrition (de perte de clientèle) en marketing, le design industriel, la gestion de stock et d'entrepôt… Ces systèmes implémentent une approche de l'IA dite symbolique. Le problème à résoudre est exprimé sous une forme symbolique à laquelle différents types de logiques peuvent être appliqués.

Leur programmation peut se faire de façon impérative ou déclarative. Dans la programmation impérative, le développeur écrit un ensemble de procédures que la machine appliquera à la façon d'une recette de cuisine ou d'une procédure de calcul. Ce niveau, perçu comme élémentaire aujourd'hui, permet néanmoins d'obtenir des résultats avec une rigueur et une rapidité inaccessibles à l'être humain. En programmation déclarative, la connaissance est décrite par un ensemble de relations entre des objets, exprimées par des formules logiques reliant des symboles. La machine peut alors indiquer les valeurs à donner à certains paramètres pour rendre possible une relation donnée. Elle est ainsi capable non seulement de calculer un résultat, mais aussi de proposer un traitement, un itinéraire, un

placement… en fonction d'un état actuel et d'objectifs donnés. De plus, la machine peut décrire le raisonnement qui a produit cette conclusion, ce qui contribue grandement à fonder la confiance dans le système.

Après avoir suscité l'enthousiasme, ces approches sont apparues plus décevantes dans les domaines tels que la traduction automatique, la reconnaissance vocale ou l'analyse d'images[1]. En effet, l'état actuel des connaissances dans ces domaines ne permet pas la création de représentations symboliques suffisamment précises pour gérer la complexité de ces tâches de façon efficace.

Depuis la victoire au jeu d'échecs de *Deep Blue* contre Kasparov en 1997, et celle au jeu de Go, d'*AlphaGo* contre Fan Hui en 2015, une autre famille de techniques occupe le devant de la scène médiatique : les méthodes statistiques dont font partie les réseaux de neurones et l'apprentissage profond ou *deep learning*. Ces méthodes tirent profit de l'augmentation de la capacité de stockage, de l'accroissement de la puissance de calcul et de la mise en réseau des ordinateurs. Elles s'appuient sur de larges ensembles de données et exploitent la connaissance implicite qu'ils contiennent sans passer par une conceptualisation ou une modélisation explicite de la connaissance. Sur la base de larges collections de données, la machine apprend à reproduire, sur une tâche donnée, des comportements similaires aux comportements les plus fréquents. Ce comportement est appris par un cerveau artificiel composé de multiples neurones artificiels organisés en couches et dont les connexions se renforcent ou s'affaiblissent au fur et à mesure des observations. Au terme de l'apprentissage, la machine est alors capable de réaliser la tâche apprise sur de nouvelles informations, elle fournit la réponse la plus vraisemblable. Ces techniques permettent de traduire un texte, sans connaître les règles de grammaire, ou d'identifier la présence d'un chat sur une photo, en ignorant même qu'un chat est vivant. Leur fonctionnement est très proche d'une boîte noire et donne très peu d'informations sur les motivations de la réponse.

1. Stevan Harnad, "The symbol grounding problem", *Physica D: Nonlinear Phenomena* 42, (1-3), 1990, p. 335-346.

Ambitions et limites induites par les caractéristiques techniques de l'IA

Sans anticiper la réflexion sur l'intelligence des techniques menée à la section suivante, une compréhension claire de leur fonctionnement permet d'identifier les points de vigilance que nécessite leur utilisation et les questions qu'elle soulève.

L'IA consiste en un ensemble d'algorithmes. Ces algorithmes, même s'ils sont capables d'apprentissage, travaillent sur des représentations du monde prédéfinies, des modèles qui sont toujours partiels. Aussi élaborée qu'elle soit, l'IA perçoit donc toujours le monde avec des œillères, elle ne prend en compte que les paramètres pour lesquels elle a été programmée. Une IA peut ainsi apprendre de façon efficace les meilleures stratégies d'un jeu en jouant des milliers de parties contre elle-même, mais toutes les parties se déroulent exactement dans le même cadre, font intervenir les mêmes éléments dans le même contexte, qui agissent selon les mêmes règles. Par contre, une démarche exploratoire qui consiste à rechercher des éléments non anticipés, à identifier les coïncidences suspectes est hors de portée d'une IA actuelle. Il est donc difficile d'imaginer une IA capable de mener des enquêtes.

Les méthodes statistiques se nourrissent de larges ensembles de données – vos données ! – et construisent des fonctions continues, apprises par interpolation à partir d'un ensemble de points discrets.

Ce que l'IA numérique apprend dans les ensembles de données, ce sont les comportements les plus fréquents, ou les plus semblables aux comportements fréquents. Ainsi, la machine a tendance à apprendre les stéréotypes. Mais corrélation ne vaut pas causalité. C'est pourquoi dans la majorité des systèmes du domaine médical, le rôle de l'IA se limite à l'aide au diagnostic, les décisions finales incombant aux médecins. Cette caractéristique suggère également la prudence qu'il faut avoir lorsqu'on considère l'intégration de tels systèmes dans le domaine de la justice qui est censée s'intéresser à l'individu dans ses spécificités. Dans le domaine peut-être moins critique du marketing, l'usage de ces techniques pour le ciblage des campagnes de promotion semble acquis. Peut-être est-il bon que seuls les nantis

reçoivent les catalogues présentant des voitures de luxe… Mais une société où seuls les nantis se voient offrir des tarifs préférentiels pour leur abonnement mobile, service de livraison à domicile et autres services peut-elle encore se dire égalitaire ?

Poussés par une logique de rendement du clic, les systèmes de recommandation qui président à la publicité ciblée sur le web, et également ceux qui définissent l'ordre de présentation des résultats dans un moteur de recherche, s'appuient sur l'exploitation des similarités : les articles qui sont proposés à l'internaute sont ceux qui ont été appréciés par des lecteurs/consommateurs ayant un profil et des habitudes semblables aux siens. Ceci peut se révéler d'une certaine efficacité et faciliter la vie. Mais, ainsi organisé par les algorithmes, notre monde a tendance à se réduire à la bulle des comportements de nos semblables, le reste du monde, tant des produits que des publications, disparaît progressivement de la portée de nos regards. C'est ainsi, par exemple, qu'après avoir consulté quelques publications complotistes ou survivalistes[2], un internaute aura rapidement le sentiment que le web est majoritairement acquis à ces thèses et attitudes.

Une fois l'apprentissage terminé, un réseau de neurones se réduit à un calculateur de cette fonction. Chaque fois qu'on lui soumettra les mêmes données, il fournira le même résultat. Si on connaît suffisamment bien la fonction, il est donc possible de forcer une réponse en manipulant les données. C'est ainsi qu'un simple autocollant fixé sur un frigo peut amener une IA à confondre ce frigo avec un grille-pain[3], de la même façon la modification d'une image d'une façon imperceptible pour l'œil humain peut amener l'identification d'un bus à une autruche[4]. Les recherches actuelles travaillent à améliorer la robustesse des solutions, néanmoins on imagine les dégâts que

2. Sont qualifiées de complotistes les personnes, publications qui défendent des thèses selon lesquelles un petit groupe de gens puissants agissant dans l'ombre seraient à l'origine des principaux événements du monde. Les survivalistes quant à eux se préparent par toute sorte de moyens à affronter un cataclysme mondial telles une guerre nucléaire ou une attaque extraterrestre.

3. Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, et Justin Gilmer, "Adversarial Patch", *31st Conference on Neural Information Processing Systems* (NIPS 2017), Long Beach, 2017.

4. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, et Rob Fergus, "Intriguing properties of neural networks", *International Conference on Learning Representations*, Banff, Canada, 2014.

pourrait provoquer un usage malveillant de cette caractéristique sur les panneaux routiers là où circuleraient des véhicules autonomes…

Un calculateur résultant d'un réseau de neurones ou d'un apprentissage profond produit une réponse à toute entrée de données, mais il n'explique pas ses conclusions. La structure des réseaux utilisés dans le *deep learning* est d'une telle dimension et complexité que leur interprétation est inaccessible à l'être humain. De recherches prometteuses[5] étudient comment améliorer l'interprétabilité des résultats en identifiant les (combinaisons d') éléments prépondérants dans la décision. Il faut noter que le choix proposé garde pour origine les corrélations présentes dans les données utilisées pour l'apprentissage, corrélations qui ne sont pas établies. Ces algorithmes sont extrêmement sensibles à la présence de biais dans les données qu'ils ne sont généralement pas capables de questionner. Ainsi, des recherches ont mis en évidence que la présence de biais racistes ou genrés dans les données d'apprentissage était reproduite par l'IA entraînée sur ces données[6].

Si l'IA de type statistique n'explique pas, c'est notamment parce qu'elle ne construit pas de modèle de connaissance ; elle exploite de façon brute la connaissance implicite présente dans les données sans la formaliser. Une telle IA peut effectuer des traductions très correctes sans connaître les principes élémentaires de la grammaire. Mais le traitement d'ambiguïtés, ou de métaphores, lui échappe. De même, les situations de dilemme ne seront abordées que par leur aspect statistique… Ainsi dans la fiction *Le Robot et le bébé* de John McCarthy[7], c'est une différence de 0.002 de valeur de

5. Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, Been Kim, "A Benchmark for Interpretability Methods in DeepNeural NetworksSara", *33rd Conference on Neural Information Processing Systems* (NeurIPS 2019), Vancouver, Canada, 2019.

6. Voir Heidi Ledford, "Millions of black people affected by racial bias in health-care algorithms", *Nature* 574 (7780), octobre 2019 ; Talia B. Gillis et Jann L. Spies, "Big Data and Discrimination", *The University of Chicago Law Review* 86, p. 459-487, 2019.

7. John McCarthy, *The Robot and the Baby*, 2004. Disponible en ligne sur http://jmc.stanford.edu/articles/robotandbaby/robotandbaby.pdf. Dans ce récit, McCarthy décrit un futur proche dans lequel un robot est chargé d'assister une maman toxicomane. Ce robot aux allures de pieuvre a été conçu pour éviter tout attachement et toute confusion affective avec un être humain, il obéit à des règles formelles. Dans le récit, une valeur de la survie de l'enfant estimée supérieure de 0.002 point à la valeur de l'obéissance à la règle de non-simulation de l'être humain amène le robot à adopter un comportement qui enfreint une série de règles en protégeant ainsi la vie du bébé. Le lecteur comprend alors qu'une infime variation de configuration aurait pu amener à privilégier l'obéissance sur la vie.

la survie du bébé qui amène le robot à enfreindre la règle lui interdisant de simuler un être humain.

Une dernière caractéristique questionne non plus l'usage de ces IA statistiques mais la collecte des données nécessaires à leur construction. De nombreuses applications, celles qui s'intéressent à un individu, qu'elles soient médicales, commerciales ou professionnelles, ont besoin de larges ensembles de données d'apprentissage issues d'individus bien caractérisés. Dans la majorité des cas, ces données sont anonymisées et servent à un apprentissage qui ne retiendra que les caractéristiques générales. Néanmoins, dans de nombreux domaines, telles les maladies orphelines par exemple, quelques caractéristiques permettent aisément de ré-identifier la personne concernée. De même, il a été établi que la connaissance de quelques clics sur les réseaux sociaux suffit pour définir assez précisément le profil de l'internaute[8]. En outre, la collecte, le stockage, le processus d'anonymisation des données se réalisent encore trop souvent à l'insu de la personne concernée. Si l'Europe cherche à protéger ses citoyens par le RGPD[9], dans de nombreux pays, les règles sont encore floues quant à qui manipule les données, qui en contrôle l'usage, qui garantit le respect la vie privée…

Intelligence et calculabilité, les limites intrinsèques de l'IA

À ce stade de l'exposé, il est utile de distinguer deux conceptions de l'intelligence artificielle, communément distinguées par les termes d'IA faible, ou spécifique, et d'IA forte, ou générale.

L'IA faible traite des systèmes qui simulent ou imitent certaines fonctions de l'intelligence humaine dans le contexte de tâches bien définies. L'IA forte ambitionne de proposer des machines très proches du raisonnement humain qui possèdent une conscience d'elle-même et des capacités d'apprentissage continu dans des environnements changeants. Si la science-fiction et les futurologues nous font parfois

8. Michal Kosinski, David Stillwell et Thore Graepel, "Private traits and attributes are predictable from digital records of human behaviour", *PNAS*, avril 2013, 110(15), p. 5802-5805.
9. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.

rêver (ou nous effraient) avec la seconde, l'état de la recherche tel que présenté dans les sections précédentes, relève clairement de la première.

Fournir des réponses intelligentes ne signifie pas être intelligent. Dans les paradigmes décrits ci-dessus, les machines raisonnent, apprennent, agissent ; et elles le font de façon efficace dans de multiples domaines : recommandation sur les sites de vente en ligne, aide à la conduite et au stationnement, identification de spam, accompagnement des entraînements sportifs et bien d'autres. Leur rapidité fascine, la pertinence de leur proposition impressionne, l'originalité surprend parfois. Toutefois, ces machines n'innovent pas vraiment, elles n'inventent pas, n'adoptent pas de comportements pour lesquels elles n'ont pas été conçues.

La proposition de recherche pour le séminaire de Dartmouth formulait l'hypothèse que « tout aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peuvent en principe être décrits avec une telle précision qu'une machine peut être fabriquée pour le simuler[10] ». Aujourd'hui, les adeptes d'une IA forte vont jusqu'à imaginer que l'on pourrait non seulement simuler mais même recréer de l'intelligence sur un autre support que le cerveau humain. Dans l'état actuel des technologies, nombre de chercheurs affichent une ambition plus modeste, limitée notamment par la théorie de la calculabilité.

Dans des travaux parallèles, Church et Turing ont étudié les fonctions qui peuvent être calculées par un algorithme, reformulé en termes d'IA, nous pourrions dire : « les comportements qui peuvent être appris par une IA ». Leurs résultats, bien connus des logiciens, sont trop nombreux et complexes pour être synthétisés ici. Toutefois, quelques idées essentielles éclairent utilement la réflexion.

Turing a proposé une formalisation d'un calculateur universel, passé dans l'histoire sous le nom de « machine de Turing ». Ce modèle abstrait d'un ordinateur est composé d'un ruban infini contenant des symboles, d'une tête de lecture/écriture sur ce ruban et d'un état interne (un entier). Un programme consiste en une table de transition : un ensemble d'instructions qui indiquent en fonction de l'état interne et du caractère sous la tête de lecture, quel caractère écrire, quel nouvel état mémoriser et quel déplacement (gauche-droite) appliquer à la tête de lecture/écriture.

---

10. John McCarthy, Marvin L. Minsky, Nathaniel Rochester et Claude E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, (traduction de l'auteur).

Les travaux des deux savants ont notamment amené à la formulation de la « Thèse[11] de Church-Turing » : « Toute fonction calculable par un algorithme est calculable par une machine de Turing[12]. » Si cette thèse n'est pas formellement vérifiée, il est toutefois établi que l'ensemble des langages de programmation développés à ce jour ont la même expressivité qu'une machine de Turing. Ceci signifie que si certains langages sont plus adaptés pour décrire certaines tâches, ils ne sont pas pour autant capables d'exprimer plus de fonctions qu'une machine de Turing.

Les machines développées avec les technologies actuelles n'ont pas d'autre intelligence que leurs algorithmes, il semble donc raisonnable d'admettre qu'elles ne peuvent adopter que des comportements calculables[13]. Bien des domaines échappent encore aux modèles : la conscience de soi, les émotions, les relations sociales… Pour les implémenter, il faudrait être capable d'exprimer ces notions en termes matériels, il faudrait que ces termes soient calculables et qu'ils puissent être transférés vers une machine… Des travaux sont en cours pour développer des machines qui analysent et simulent des émotions[14]. Certains auteurs et chercheurs rêvent de machines qui éprouveraient des émotions, mais la technologie qui permettrait de les développer reste à découvrir.

La compréhension des mécanismes de l'intelligence humaine n'est, elle aussi, encore que très partielle. Non seulement nous sommes encore loin de pouvoir répliquer le fonctionnement du cerveau, mais, de plus, il est maintenant reconnu que la pensée et l'intelligence humaines dépassent de loin le cadre strict du cerveau[15].

---

11. Une hypothèse non prouvée, mais jusqu'ici jamais falsifiée et donc communément admise.
12. Voir le chapitre, « La thèse de Church », dans l'ouvrage de Gilles Dowek, *Les métamorphoses du calcul. Une étonnante histoire de mathématiques*, Paris, Le Pommier, 2011 (2007), p. 81-100 ; Maël Pégny, « Les deux formes de la thèse de Church-Turing et l'épistémologie du calcul », *Philosophia Scientiae* 16(3), 2012, p. 39-67.
13. Dans cet article, nous ne discutons pas la distinction entre les deux formulations de la thèse de Turing (algorithme ou machine). Pour plus de détails sur cette distinction et ses implications sur l'épistémologie du calcul, le lecteur se référera à Maël Pégny, « Les deux formes de la thèse de Church-Turing et l'épistémologie du calcul ». Bien que l'équivalence n'aie pas été établie, nous n'envisageons ici que les machines qui exécutent des algorithmes, et n'entrons pas dans le débat.
14. Thomas Moerland, Joost Broekens et Catholijn M. Jonker, "Emotion in reinforcement learning agents and robots: a survey", *Machine Learning*, 2018, n° 107, p. 443-480.
15. Miguel Benasayag, « La pensée n'est pas dans le cerveau ! », propos recueillis par R. Meyran dans *Le Courrier de l'Unesco*, juillet-septembre 2018.

Les travaux de Turing sur les capacités de la machine l'ont également amené à questionner la notion d'intelligence et à proposer le fameux « test de Turing ». Il suggère qu'une machine pourra être jugée intelligente si elle est capable de dialoguer avec un être humain de façon telle que, les modalités d'interaction étant dissimulées, l'interlocuteur ne puisse pas discerner s'il interagit avec une machine ou un être humain[16]. La pertinence et les modalités du test font l'objet de nombreux débats. Néanmoins, ce test fait l'objet d'un concours annuel, le prix Hugh Loebner, et les résultats des machines sont en progrès constants.

Récemment, les progrès dans la simulation des compétences relationnelles et émotionnelles des agents sociaux ont suscité une proposition d'adaptation du test de Turing : le « gift test ». Ce test consiste à évaluer la capacité de la machine à faire plaisir à un être humain dans la durée.

La liste n'est pas encore épuisée des capacités pas ou mal simulées par l'IA : intérêt, émotions, sens des responsabilités, créativité, pertinence… Certains chercheurs[17] suggèrent que la mesure de la simplicité, telle que modélisée par des techniques issues de la théorie de l'information[18], pourrait ouvrir des voies de progression dans le développement de ces compétences. L'identification d'un objet ayant une simplicité atypique dans leur ensemble est en effet un déclencheur d'intérêt voire d'émotion pour l'être humain.

De la responsabilité et de la moralité de l'IA

La présence d'IA de plus en plus autonomes, embarquées ou non dans des robots, est déjà très large, et ne cessera d'augmenter. Certaines applications techniquement avancées posent néanmoins une série de questions quant à leur intégration dans notre quotidien, dans

16. Alan Turing, "Computing Machinery and Intelligence", *Mind*, vol. LIX, n° 236, octobre 1950, p. 433-460.

17. Jean-Louis Dessalles, *Des Intelligences TRÈS Artificielles*, Odile Jacob, 2019.

18. Kolmogorov associe la complexité d'un objet au nombre d'informations de la description la plus synthétique qui peut en être faite, ainsi la suite « aaaaaa » est considérée comme plus simple que « lphdfs ».

la société. Une IA peut-elle être responsable, morale, éthique[19] ? Ces questions sont largement transdisciplinaires et demanderaient d'être étudiées dans leurs composantes robotique, juridique, sociologique, psychologique, philosophique, éthique et morale. Je propose d'évoquer ici quelques éléments de réflexion apportés par la perspective informatique.

Tant que l'IA est confinée dans des outils d'aide à la décision, le décideur humain garde la main sur la décision, et il en porte la responsabilité. Les agents, bots[20] et robots devenant de plus en plus autonomes, la question de la responsabilité de leur comportement devient complexe, d'autant plus complexe que le niveau d'autonomie se décline en une large gamme de scénarios : IA supervisée, IA autonome sur des tâches prescrites, IA totalement autonome et apprenante, et même IA qui supervise l'homme, avec ou non, pour chaque niveau, la possibilité de basculer entre ces modes[21].

Comme dans d'autres domaines, des processus de certification par un tiers de confiance peuvent contribuer à articuler la relation entre concepteur, producteur et acquéreur. Mais les choses deviennent nettement plus complexes lorsque l'agent commercialisé peut être supervisé, voire paramétré ou partiellement (re)programmé ; plus complexes encore lorsqu'on parlera d'agent capable d'apprendre des tâches nouvelles. Le propriétaire, les utilisateurs sont alors eux aussi impliqués dans le comportement de la machine. Certains auteurs proposent de résoudre ce nœud de responsabilités au moyen d'une fiction juridique en faisant de la machine un agent porteur de sa propre « responsabilité », mais ceci ne semble pas apporter une réponse satisfaisante dans tous les scénarios[22].

La question de la morale des machines nous ramène à la nature de ce qu'une machine peut traiter ; dit simplement : des fonctions ;

19. Le débat sur ce qui relève spécifiquement de la morale ou spécifiquement de l'éthique dans ce qui suit nous emmènerait au-delà du propos de cet article. Comme il s'agit ici de concepts généraux, même s'ils sont illustrés dans des contextes spécifiques, j'utilise par la suite uniquement le terme « morale », de façon peut-être un peu abusive.

20. Un bot est un programme qui s'exécute de façon autonome dans un univers digital, il se distingue ainsi du robot par le fait qu'il ne possède pas de « corps » pour interagir avec le monde réel.

21. Dominique Lambert, *Éthique et Robotique, quel pouvoir peut-on laisser aux machines ?* Communication orale lors des Grandes Questions de la Philosophie, Séminaire du département de Philosophie, Namur, 18 avril 2017.

22. Nathalie Nevejan, *Traité de droit et d'éthique de la robotique civile*, LEH Édition, 2017.

de façon plus élaborée : des connaissances complètement exprimables de façon formelle ou numérique. Quelle forme de morale pourrait donc être implémentée par une machine ? Les contraintes de l'implémentation porteront inévitablement à se tourner vers une morale de type utilitariste. Si même il était acquis que ce type de morale soit une bonne solution, et pas seulement un choix par défaut, nombre de questions restent ouvertes : quelles règles faut-il implémenter ? Et préalablement, qui fixe les règles à implémenter ? Le concepteur, l'utilisateur, l'État… ? Resterait alors encore à trancher comment traiter les dilemmes, tout en restant conscient que, avec les technologies actuelles, la machine ne prendra en compte aucun élément qui n'ait été introduit dans son modèle du monde, quelle qu'en soit la pertinence.

Des recherches tentent de proposer des modes de programmation qui assurent que les IA soient « éthiques par leur conception » (*ethical by design*). Des premiers résultats, prometteurs, visent à assurer des qualités telles que l'exécution correcte des algorithmes, la confidentialité des données, l'identification et l'élimination de biais dans les données… Mais les outils de formalisation nécessaires à la traduction de concepts éthiques de haut niveau contraignant le code semblent encore inaccessibles. Des lois du type de celles d'Isaac Asimov[23] sont encore loin de pouvoir être implémentées de façon robuste.

Compte tenu de ces caractéristiques, la pertinence et l'opportunité d'utiliser l'IA dans divers domaines demande une analyse fine de ces domaines : jusqu'à quel niveau d'autonomie et dans quelles limites une IA peut-elle soigner, opérer, investir ? Pourrait-elle juger ? Pourrait-elle tuer et faire la guerre ?

L'IA DANS LA SOCIÉTÉ

L'intégration des IA dans la société a commencé, poussée par les industries et l'innovation technologique, de nombreuses autres

23. Isaac Asimov, "Runaround", *Astounding Science Fiction*, 29(1), 1942, p. 94-103. (1. Un robot ne peut porter atteinte à un être humain, ni, en restant passif, permettre qu'un être humain soit exposé au danger ; 2. Un robot doit obéir aux ordres qui lui sont donnés par un humain, sauf si de tels ordres entrent en conflit avec la première loi ; 3. Un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou deuxième loi.) (NDLR).

applications verront le jour dans les prochaines années, et elles pourront avoir des conséquences significatives sur nos modes de vie, de relation, de travail. Actuellement, les balises sont faibles et demeurent très informelles quant aux règles à respecter pour ces développements. La séduction exercée par le progrès guide trop souvent les choix. L'intégration de l'IA dans notre société ne manque pas d'interroger l'avenir, nos projets et la nature même de l'homme. Les machines pourraient-elles nous diriger un jour ? La distinction entre l'homme et la machine pourrait-elle s'estomper ?

Dans l'état actuel des technologies, la crainte que les machines prennent le pouvoir semble peu fondée. Mais il est une prise de pourvoir plus invisible qui pourrait bien être en marche. L'introduction de l'IA dans de multiples domaines pourrait bien modifier nos façons d'analyser les questions en les réduisant à des formes évaluables par une IA. Nombre de systèmes d'investissement, de sélection de candidat, de gestion de carrière ont déjà pris ce chemin, soutenus par le sentiment que le quantifiable est seul objectivable et juste, sans en questionner l'adéquation. Or, la logique reconnaît l'incomplétude, ce que les mathématiques prouvent est vrai, mais elles n'épuisent pas le réel. L'anthropologie souligne que l'humain ne se réduit pas à des nombres… L'objectivité mathématique cache trop souvent l'incapacité à faire confiance, à prendre un risque, or c'est de la créativité, du « non-standard » que peut jaillir la plus grande fécondité[24].

Par ailleurs, la logique de la rentabilité maximale, l'instantanéité de la mise à disposition des informations et de leur traitement induisent déjà une forme d'asservissement au rythme des machines, une disparition progressive du temps long de la réflexion, de la maturation, de la créativité.

Des addictions nouvelles apparaissent notamment liées aux mondes virtuels. Il est déjà douloureux pour beaucoup de vivre la séparation d'un objet, d'une machine sans intelligence qui a accompagné une part de leur chemin et rendu de précieux services : montre, voiture, maison, outil professionnel… ou simple briquet ! Combien plus fort sera cet attachement quand, de plus, ces objets devenus « intelligents » donneront de façon crédible l'illusion de vie, d'émotions, de sentiments ?

24. Dominique Lambert, *La Robotique et l'Intelligence Artificielle*, Fidélité, 2019.

L'IA apporte des contributions significatives aux traitements de multiples pathologies, à la réalisation de prothèses de plus en plus performantes. Fort de ces résultats, le projet transhumaniste imagine d'exporter ces résultats au-delà du champ thérapeutique, et explore les possibilités d'un homme augmenté, intégrant dans son propre corps, jusque dans son cerveau, des composants artificiels, voire de transférer ses compétences cognitives dans un corps synthétique[25]. Beaucoup de ces ambitions tiennent encore aujourd'hui de la science-fiction, néanmoins elles nous interrogent sur la possible confusion entre l'homme et la machine. Des résultats, notamment sur l'usage des GPS par les chauffeurs de taxi[26], mettent en évidence qu'augmenté artificiellement, l'homme peut s'en trouver diminué dans ses capacités propres. Par ailleurs, psychologie et anthropologie nous indiquent que le corps ne peut pas être pensé comme un simple contenant, mais est une part intégrante de la personnalité.

Sensibles à ces questions, on ne peut s'empêcher d'en entendre surgir d'autres. Qui mène la recherche en IA ? Qui en a les moyens[27] ? À quels problèmes l'IA s'intéresse-t-elle ? Seulement ceux de l'homme blanc[28] ? Qui développe les programmes de financement ? Jusqu'à quel point est-on obligé d'utiliser l'IA ? Jusqu'à quel point est-on libre de refuser d'utiliser l'IA ? Autant de questions auxquelles l'informaticien ne peut répondre seul.

La nécessité d'un dialogue avec les sciences humaines

Le développement de technologies nouvelles s'accompagne du devoir d'éduquer, de penser un projet, de construire de bonnes pratiques et de les formaliser dans des règles et des lois là où cela s'impose ; d'autant plus quand les usages de ces technologies ne se

25. Sur ce sujet, voir le dossier de la *RETM* 302, juin 2019 : *Le transhumanisme : une religion ?* (NDLR).

26. Voir Miguel Benasayag, « La pensée n'est pas dans le cerveau ! ».

27. Yeshua Bengio, « Contrer la monopolisation de la recherche », propos recueillis par J. Šopova dans *Le Courrier de l'Unesco*, juillet-septembre 2018.

28. Moustapha Cissé, « Démocratiser l'IA en Afrique », propos recueillis par K. Markelova, dans *Le Courrier de l'Unesco*, juillet-septembre 2018.

limitent pas à un domaine spécifique mais sont disséminés dans de multiples domaines de la vie.

Du côté de la réflexion menée par les développeurs, il est intéressant de souligner l'initiative *Ethically Aligned Design* de l'IEEE, société savante visant notamment à la définition de standards dans l'industrie. Au terme d'un long processus de collaborations internationales, ce travail a produit en janvier 2020 une liste de recommandations pour l'évaluation de l'impact des systèmes autonomes et intelligents sur le bien-être humain[29]. Ces recommandations prennent en compte une vision holistique s'appuyant sur la triple perspective des personnes, de la planète et du profit en vue d'accorder la priorité à une vision à long terme du bien-être humain tant sur le plan de la santé physique que mentale.

Les questions posées par l'intelligence artificielle vont bien au-delà des champs de l'ingénieur, du roboticien et de l'informaticien, elles demandent un dialogue avec bien d'autres disciplines : neurologie, psychologie, droit, sociologie, anthropologie et éthique. Des travaux approfondis ont été déjà réalisés par des comités d'éthique, notamment le Comest de l'Unesco[30] et, en France, la Cerna[31] ; tous deux ont publié leur rapport en 2017.

Dans cet article, j'ai présenté ce qui peut être dit de l'intelligence artificielle depuis la perspective d'un informaticien, j'ai tenté d'identifier les caractéristiques qui me semblent ouvrir des questions quant aux attentes qu'elles peuvent ou non rencontrer ainsi que celles concernant leurs modalités d'intégration dans divers domaines. J'espère que les chercheurs d'autres disciplines y trouveront des fondements utiles à un dialogue fécond au sujet de l'intelligence artificielle, sa nature et ses possibles usages.

Isabelle Linden
Université de Namur,
Namur Digital Institute (NADI)

29. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition, IEEE, 2019.

30. Rapport de la COMEST sur *L'Éthique de la Robotique*, UNESCO, SHS/YES/COMEST-10/17/2 REV, 2017.

31. CERNA Allistene, rapport *Éthique de la recherche en apprentissage machine,* édition provisoire, juin 2017.