Quelques questions éthiques autour du développement de l'autonomie décisionnelle en intelligence artificielle et en robotique

Mehdi Khamassi

Quelques questions éthiques autour du développement de l'autonomie décisionnelle en intelligence artificielle et en robotique

MEHDI KHAMASSI

Mehdi Khamassi est directeur de recherche en sciences cognitives au CNRS, rattaché à l'Institut des systèmes intelligents et de robotique de Sorbonne Université, Paris. Il est également co-directeur des études pour le master de sciences cognitives de l'École normale supérieure / EHESS / Université de Paris. Après un diplôme d'ingénieur à l'ENSIIE (anciennement sous la tutelle du CNAM) à Evry, il a effectué un DEA de sciences cognitives (Cogmaster) puis une thèse entre l'Université Pierre et Marie Curie et le Collège de France dirigée par Agnès Guillot et Sidney I. Wiener sur l'apprentissage en situation de navigation chez les animaux et les robots.

Le terme « Intelligence Artificielle » (IA) désigne au départ un ensemble de programmes informatiques qui résolvent des problèmes habituellement résolus par des processus mentaux de haut niveau chez l'humain. Cette définition a émergé d'un projet pour un workshop à Dartmouth College en 1956, projet auquel ont notamment participé les chercheurs John McCarthy, Marvin Minsky, Herbert Simon, Alan Newell, Nathaniel Rochester et Claude Shannon.

Dès le départ, dans le projet de l'IA, l'objectif était aussi de comprendre ces processus cognitifs de haut niveau chez l'humain, en cherchant à les modéliser puis à les simuler. Cela a souvent donné lieu à un découpage du travail et à une répartition entre équipes de chercheurs selon les différentes fonctions cognitives : le langage, la perception, le raisonnement, avec l'idée de réunir ensemble ces éléments, une fois chacun achevé, pour créer un agent intelligent. Au cours des années 1970, une forme de critique émerge, avec le constat des limites de ce type d'approche. D'une part, modéliser chaque fonction de manière isolée et ensuite espérer « coller » les morceaux, faire que cela fonctionne en connectant simplement les entrées et les sorties des fonctions entre elles, était bien illusoire. En neurosciences aujourd'hui, on se rend compte que les réseaux cérébraux impliqués dans différentes fonctions cognitives sont beaucoup plus distribués et étendus dans le cerveau qu'on ne le pensait avant. De nombreuses aires cérébrales participent à plusieurs fonctions cognitives différentes. Une deuxième forme de critique émerge pour pointer les limites de l'IA symbolique : le fait de décrire le problème par un ensemble de symboles discrets, chacun décrivant une variable isolée du

problème, et ensuite faire du raisonnement logique de premier ordre sur ces symboles. En effet, une des faiblesses de cette approche a été clairement illustrée lors de la confrontation d'agents artificiels au monde réel, comme des robots par exemple. Dans ce cas, il s'avère que la variabilité des signaux mesurés par les capteurs du robot, l'incertitude liée aux mesures, et la diversité des situations rencontrées obligent à décrire une tâche donnée avec un très grand nombre de symboles. Ceci représente une explosion combinatoire qui se produit quand chaque symbole décrit une configuration précise de multiples variables : il existe un très grand nombre de combinaisons possibles de ces variables qu'il faut traiter. Exemple : une boîte est présente sur la table, mais cette boîte est ouverte (donc le robot devra la refermer avant de la saisir), et le robot a déjà une boîte dans un de ses bras, or il s'agit d'une grosse boîte qui a besoin d'être saisie avec les deux bras, etc. De plus, si un symbole représente une configuration très précise de la tâche, alors ce symbole ne pourra pas s'appliquer à d'autres tâches légèrement différentes, conduisant ainsi à une incapacité du programme informatique à généraliser. C'est donc à ce moment-là que l'on voit émerger d'autres approches de l'IA, notamment des approches connexionnistes, pour permettre un calcul davantage distribué, dont certaines neuro-inspirées, des approches probabilistes, etc.

Pour affiner la définition de l'IA et mieux appréhender certains enjeux d'aujourd'hui, il faut d'abord revenir sur la distinction entre ce qu'on appelle "l'IA faible" et "l'IA forte". L'IA faible relève d'un domaine de recherche de l'Intelligence artificielle qui ne cherche pas (ou plus) à produire un algorithme *intelligent*, au sens de l'intelligence humaine, mais plutôt à trouver des méthodes algorithmiques avancées et efficaces pour extraire de la connaissance à partir des données. Il s'agit donc d'un champ de recherche proche de la science des données et de la statistique. C'est dans ce domaine que s'épanouit ce qu'on appelle l'apprentissage automatique, ou l'apprentissage machine, *machine learning* en anglais. De l'autre côté, les recherches en IA forte continuent de viser à produire des capacités intellectuelles proches de celles de l'humain, avec parfois la quête d'agents artificiels conscients ou capables de métacognition. À ces fins, il est important de considérer la manière dont différentes fonctions cognitives sont intriquées, communiquent au travers de l'interaction entre des systèmes multiples de mémoire (mémoire de travail, mémoire épisodique, mémoire procédurale, mémoire à long-terme). Ceci est mis en œuvre le plus souvent dans ce qu'on appelle des « architectures cognitives », qui proposent des hypothèses sur la manière d'articuler ces différents processus.

Une autre distinction importante et intriquée avec la précédente est celle entre la robotique et la modélisation cognitive. Cette dernière a pour objectif de modéliser des phénomènes qui existent chez l'humain (ou chez l'animal), avec une focalisation sur des facultés cognitives spécifiques, par exemple l'apprentissage ou la perception. La vision holistique ne permet (pour l'instant) pas vraiment de modéliser et de comprendre finement des phénomènes donnés et d'exploiter des données expérimentales. Les approches systèmes ou *whole-brain* en modélisation pour les neurosciences ont donc été partiellement mises de côté depuis environ 20 ans pour laisser place à la modélisation plus précise et plus parcimonieuse (i.e., avec moins de paramètres) de fonctions isolées. Mais le besoin se fait de plus en plus sentir en neurosciences computationnelles de revenir à des approches systèmes, ou tout du moins, d'alterner entre modèles de fonctions très spécifiques, et modèles de grands réseaux cérébraux faisant cohabiter plusieurs fonctions cognitives différentes. Parfois, même le rôle du corps dans la cognition, donc la dimension incarnée, s'avère nécessaire à prendre en compte dans les modèle : par exemple, lorsqu'un chercheur comme Kevin O'Regan nous dit que percevoir une éponge devant nous, c'est en partie ressentir à nouveau la sensation de l'interaction physique dont nous avons fait précédemment l'ex-

périence lorsque nous pressions une éponge dans notre main. C'est suivant cette évolution que la modélisation cognitive interagit de plus en plus avec les recherches en robotiques cognitives, qui font intervenir le corps du robot dans la résolution d'une tâche, contrairement à l'IA qui est le plus souvent simulée, donc désincarnée. En ce qui concerne la robotique, pour reprendre une définition de Jean-Paul Laumond[1] : un robot est un ordinateur doué de mouvement, en d'autres termes un agent physique qui se déplace dans l'espace et interagit ainsi avec l'environnement. À la fin des années 1960, une grande partie des recherches en robotique reposait sur l'IA symbolique. Ceci peut être illustré par les travaux sur le robot Shakey, menés au Stanford Research Institute, dans lesquels des algorithmes d'IA symbolique manipulent des symboles pour vérifier un certain nombre de conditions permettant au robot un plan d'actions conduisant à des déplacements dans l'espace. Aujourd'hui, la robotique est divisée en nombreuses sous-disciplines. Deux d'entre elles, sur lesquelles je vais m'attarder, sont particulièrement pertinentes par rapport à l'autonomie décisionnelle : 1) la robotique développementale, qui cherche à mimer un processus d'apprentissage progressif du robot en interaction avec le monde, comparable à celui qui est développé par l'enfant humain. 2) la neuro-robotique, qui cherche à tester des solutions algorithmiques pour mimer des processus neurobiologiques. Dans les deux cas, on trouve à la fois l'inspiration de la biologie et l'objectif de mieux comprendre des processus biologiques en testant des hypothèses sur des robots. Ces deux sous-disciplines entrent dans le cadre de la robotique autonome qui nécessite de doter le robot d'un certain niveau d'autonomie décisionnelle et de capacités d'adaptation.

L'autonomie décisionnelle est une expression très employée en robotique dans un champ qui fait souvent référence à des travaux de philosophie et de psychologie pour discuter de l'autonomie éventuelle des systèmes artificiels. Pour clarifier cette notion d'autonomie, on peut faire référence[2] à McFarland (1995), qui la définit ainsi : « L'autonomie implique la liberté par rapport à un possible contrôle externe [...]. Un agent autonome doit avoir un certain degré de motivation et de cognition organisés d'une façon qu'un agent extérieur ne puisse pas obtenir suffisamment d'information pour contrôler l'agent autonome. ».

À l'idée, mise en évidence dans ce passage, qu'un agent autonome ne peut pas être complètement déterminé par autrui, on peut attribuer comme cause non seulement le fait qu'un agent externe n'a pas assez de connaissance pour le contrôler mais aussi que l'agent externe ne peut pas parfaitement *prédire* le comportement d'un agent autonome.

Un autre argument pour élargir le concept d'autonomie est celui de l'autosuffisance. Pour être autosuffisant, un robot devrait pouvoir fonctionner sur le long terme et pas seulement pendant le temps d'exécution d'une tâche ; en d'autres termes, il devrait en permanence entretenir son niveau d'énergie. À ce propos, McFarland souligne que, si en plus d'être autonome, le robot est aussi autosuffisant, alors cela inclut « la motivation appropriée [...] pour décider quand se recharger en énergie, pour éviter les températures extrêmes, pour éviter les prédateurs et les pièges, et pour effectuer ses tâches. » Il est question ici de la capacité non seulement de décider de façon autonome des actions ponctuelles commandées par l'humain mais aussi de décider de réaliser une série d'actions pour répondre aux propres be-

[1] Jean-Paul Laumond, « Robotique : l'intelligence de la gravité », *Les Cahiers de TESaCo* N°1 (2021).

[2] David McFarland, Opportunity versus goals in robots, animals, and people, *in* H.L. Roitblat & J.-A. Meyer (eds.), *Comparative Approaches ot Cognitive Science*, Cambridge, MA : MIT Press, 1995, p. 416. Nous avons abordé ces sujets dans une synthèse des travaux en robotique cognitive écrite avec Stéphane Doncieux : Khamassi, M., & Doncieux, S. (2016). Nouvelles approches en robotique cognitive. *Intellectica*, 65(1), 7-25.

soins de l'agent de se maintenir dans un état viable. En psychologie cognitive, on retrouve cette idée sous les termes d'« autonomie psychologique », d'« autonomie individuelle » ou simplement d'« autonomie ». Des travaux en psychologie, comme par exemple l'article récent *From freedom from to freedom to. New perspectives on intentional action* (Bonicalzi et Haggard 2019)[3] sont en résonance avec les études en neuro-robotique et en robotique développementale dans lesquelles il s'agit d'utiliser le robot comme outil de modélisation pour comprendre des phénomènes qui ont lieu chez l'humain. Dans l'article cité, on lit : « Le concept d'avoir le choix est central pour la discussion métaphysique sur le libre arbitre et le déterminisme » (van Inwagen, 1983; Pereboom, 2014). Ceci est abordé la plupart du temps sous forme de la question de savoir si les individus ont la possibilité d'agir autrement, et ont donc le choix, lorsqu'ils décident de ce qu'ils vont faire. Dans cet article, nous mettons de côté la question de savoir si les gens ont le choix au sens métaphysique. Nous adoptons à la place un concept plus limité de la notion de choix : pour tout ce qui touche au comportement planifié (e.g., Pierre décide au temps t0 s'il va aller à Paris au temps t1), les individus ont la faculté de s'engager dans un processus de raisonnement orienté vers un but, d'exprimer leurs préférences à propos d'options apparemment disponibles, et finalement d'agir selon ces préférences. La notion d'autonomie psychologique d'Alfred Mele est parfaitement en accord avec ces desiderata. Selon Mele[4], une version compatibiliste de l'autonomie psychologique est satisfaite quand les trois conditions suivantes tiennent conjointement : (1) l'agent n'est pas sous l'influence d'états motivationnels coercitifs ou convaincants ; (2) les croyances de l'agent sont propices à une délibération éclairée ; (3) l'agent est un délibérateur fiable (Mele, 1995).

La recherche en robotique sur l'autonomie décisionnelle

Un certain nombre de recherches en robotique étudient comment doter les robots d'autonomie, en les rendant capables de planifier leurs suites d'actions pour atteindre un but précis. Il reste pourtant que l'état motivationnel, le but de l'agent, a été déterminé de façon externe par l'humain, de sorte que l'on peut toujours mettre en doute ou relativiser cette autonomie. Néanmoins, dans certains travaux, on peut définir une règle générale de maintien de l'état du robot (maintien du niveau d'énergie tout en accomplissant ses tâches) au sein de laquelle le robot est autonome ensuite pour décider à quel moment il travaille sur ses tâches et à quel moment il va se recharger en énergie ou satisfaire d'autres objectifs (interaction sociale, quête d'information sur le monde, ce qu'on appelle la « curiosité artificielle », etc.). Or, ces travaux emploient des modèles de prise de décision qui ne sont parfois pas très éloignés sur les principes des modèles de prise de décision chez l'humain, tels qu'établis en psychologie cognitive. Donc, la recherche de modèles conférant une autonomie satisfaisante chez un robot peut éclairer la recherche de la définition de l'autonomie chez l'humain. Il s'agit dès lors de mieux définir, de manière philosophique, la notion d'autonomie, par exemple en la comparant avec celle de libre arbitre et d'autonomie psychologique et décisionnelle chez l'humain[5]. Sur ce sujet, un article de Daniel Andler (responsable de TESaCo), issu du numéro spécial d'*Intellectica* "Éthique et sciences cognitives" que j'ai coordonné en 2019 avec Alain Mille et Raja Chatila, traite de la spécificité de certaines questions éthiques en sciences cognitives, et

[3] Bonicalzi, S & Haggard, P (2019). From *Freedom From* to *Freedom To* : New Perspectives on Intentional Action. *De face. Psychol* . 10: 1193. doi: 10.3389 / fpsyg.2019.01193

[4] Alfred R. Mele, Autonomous Agents: From Self Control to Autonomy, *Oxford University Press* (1995).

[5] Voir Atlan, H. (2018). *Cours de philosophie biologique et cognitiviste: Spinoza et la biologie actuelle.* Odile Jacob.

aborde aussi la question du libre arbitre, et donc du degré d'autonomie dont disposent les humains pour prendre leurs décisions. L'article souligne notamment que les recherches en sciences cognitives ont une responsabilité particulière du fait que certaines des conclusions (provisoires) ou interprétations de ces travaux, notamment sur le libre arbitre, peuvent « modifier[…] notre conception de l'humanité en nous », et donc avoir des conséquences sociétales importantes. Par conséquent, il est important que les chercheur.e.s en sciences cognitives prennent des précautions sur la façon dont certains messages, parfois trop concis ou trop tranchés, peuvent être interprétés dans la société, alors que ces informations scientifiques s'appliquent dans le cadre limité et restreint du laboratoire sous des conditions particulières. Dans la même veine, un article récent d'Ariane Bigenwald (qui a participé au comité scientifique de TESaCo), co-écrit avec Valérian Chambon, traite de la question du possible impact sur le droit des réflexions en sciences cognitives (et en particulier en neurosciences) concernant l'absence de réel libre arbitre, en défendant l'idée que celui-ci constitue une notion bien séparée de celle de responsabilité pénale[6] utilisée dans l'institution judiciaire.

Questions éthiques de l'IA

Plusieurs problèmes éthiques sont liés en général à l'IA. Parmi ceux-ci : 1) l'efficacité de l'IA faible : le processus d'extraction des connaissances à partir des données repose sur le croisement d'une grande quantité de données, souvent acquises de manière commerciale, ce qui pose un certain nombre de questions éthiques (accès au données, croisement des données, partage des données, anonymisation) ; 2) la possibilité, désormais effective, comme le montre l'exemple du scandale Cambridge Analytica, d'utiliser l'IA faible à des fins d'influence (marketing, publicité, campagnes politiques, etc.) ; 3) l'absence de transparence par rapport à quelles données personnelles sont traitées par les algorithmes ; 4) un manque de transparence également sur la nature, les opérations et les finalités des algorithmes avec lesquels l'humain interagit.

Quelle est la pertinence de ces problèmes pour l'IA et la robotique ? La recherche en IA forte et en robotique autonome est consacrée aux principes par lesquels un agent artificiel peut atteindre un certain degré d'autonomie décisionnelle. Or, le développement d'une autonomie en robotique peut produire des conséquences sociétales. Comme évoqué plus haut, les principes à la base de l'autonomie sont souvent étudiés, d'un côté, dans le but de les comprendre au niveau computationnel pour ensuite, sur cette base, programmer la machine, de l'autre pour modéliser et mieux saisir le processus sous-jacent chez l'humain. L'autonomie développée chez les robots présente aussi des aspects applicatifs, car de nombreuses applications de robots nécessitent aussi un certain degré d'autonomie. Il s'agit, par exemple, de l'usage des robots là où l'humain ne peut pas intervenir (exploration de Mars ou des fonds marins, intervention dans une zone radioactive, etc.) ou dans des situations pour lesquelles on a raison de penser que l'humain interviendrait de façon plus lente et moins efficace que le robot, ou que l'algorithme (par exemple les ordinateurs utilisés pour le pilotage automatique des avions).

Avant de passer aux potentiels impacts et enjeux éthiques de ces technologies, je vais présenter deux travaux permettant d'illustrer plus précisément en quoi consiste le développement de l'autonomie décisionnelle chez les robots. Dans l'étude Chatila et al. (2018)[7], une expérience en neuro-robotique est décrite : un

[6] Bigenwald, A., & Chambon, V. (2019). Criminal responsibility and neuroscience: no revolution yet. *Frontiers in psychology*, 10, 1406.

[7] Chatila, R., Renaudo, E., Andries, M., Chavez-Garcia, R. O., Luce-Vayrac, P., Gottstein, R., Alami, R., Clodic, A., Devin, S., Girard, B. & Khamassi, M. (2018). Toward self-aware robots. *Frontiers in Robotics and AI*, 5, 88.

robot naviguant dans l'espace rencontre des obstacles et, au fur et à mesure qu'il explore son environnement, il construit une carte suivant certains principes, par exemple comment se déplacer de manière efficace pour ne pas toujours revenir aléatoirement au même endroit. C'est le robot qui découpe de manière autonome l'espace en lieux discrets qui vont lui permettre de réaliser une planification plus efficace. Le robot fait un apprentissage par essais et erreurs dans lequel il va apprendre comment choisir les bonnes actions : étant donné les huit directions cardinales, il va décider dans chaque position quelle est la meilleure action à mettre en place pour rejoindre une certaine zone, but qui est déterminé par l'expérimentateur. Dans cette expérience, assez représentative de nombreux travaux conduits en robotique, on dote le robot d'un programme informatique appelé en robotique une « architecture de contrôle » ou une « architecture cognitive » : celle-ci organise différentes couches informationnelles au sein du robot, dont par exemple la couche décisionnelle. Les couches supérieures envoient des ordres d'actions vers le bas, voire vers des couches exécutives qui vérifient que tout soit bien en place pour réaliser l'action prévue. Trois boîtes sont présentes à l'intérieur de la couche décisionnelle : elles correspondent aux différentes stratégies de prise de décision et d'apprentissage sur la base des conséquences de l'action. En même temps, un méta-contrôleur observe comment le robot apprend et arbitre parmi les stratégies utilisées. Le robot va ainsi, petit à petit, apprendre quelle est la bonne stratégie d'apprentissage. La question qui se pose à partir de cet exemple est la suivante : quel degré d'autonomie est développé dans ces travaux en robotique ? Le fait qu'il y ait un choix autonome de l'action (i.e. dans quelle direction se déplacer et quelle stratégie d'apprentissage choisir) constitue certes un certain degré d'autonomie. Deuxièmement, la présence d'une sorte de métacognition élémentaire (le fait d'apprendre que, dans une situation, la stratégie A a été plus efficace que la stratégie B) témoigne aussi de la présence d'une certaine autonomie. En effet, le robot va pouvoir déterminer tout seul, après apprentissage, sa préférence pour utiliser la stratégie A dans tel et tel contexte, sans que l'humain intervienne. Pourtant, dans ces expériences, le but est toujours fixé par l'humain, si bien qu'il est déterminé par un agent extérieur : de ce point de vue, il n'y a pas encore d'autonomie telle que je l'ai définie au départ par les mots de McFarland. En l'occurrence, on considère qu'il n'y a pas de motivation intrinsèque au robot dans la décision d'atteindre un lieu ou un but plutôt que de réaliser une autre tâche : il n'y a donc pas d'autonomie en ce sens. Si l'on fait le lien entre ce modèle et celui de la prise de décision chez l'humain, on se rend compte qu'il existe une grande quantité de points de convergence. La prise de décision chez l'humain contient tous ces ingrédients. Chez l'humain, on trouve pourtant un élément supplémentaire : une articulation de motivations extrinsèques et intrinsèques. Pour illustrer cette idée, imaginons un espace à plusieurs dimensions dont chacune correspond à une motivation (par exemple, maintenir sa température du corps dans une zone de viabilité ou maintenir son niveau de glucose en-dessous d'un certain seuil de viabilité). Certaines de ces motivations sont extrinsèques, telle remplir une tâche donnée, ou se recharger en énergie. D'autres sont des motivations intrinsèques, comme celle de vouloir acquérir de l'information à propos d'un environnement, exprimée par un certain niveau de curiosité. Dans les propositions les plus récentes, le modèle de prise de décision chez l'humain est caractérisé, d'une part, par le simple fait de combiner ces différentes motivations et, d'autre part, par le fait que l'algorithme va ensuite essayer de se maintenir dans une zone de viabilité dans cet espace (dont les dimensions peuvent être multiples) : si le ni-

veau d'énergie diminue, l'agent va être amené à accomplir une tâche qui lui redonnera de l'énergie pour maintenir son équilibre homéostatique. Dès que ces deux éléments extrinsèques et intrinsèques sont combinés, les modèles semblent avoir un fort pouvoir explicatif de prise de décision chez l'humain, du moins dans un cadre de laboratoire (voir par exemple Friston et al., 2015)[8].

Je vais à présent citer un dernier exemple de travaux[9] en robotique développementale dans lesquels le principe de motivation intrinsèque est mis en avant (voir *The playground Experiment*, dans Oudeyer, 2018). Cette expérience utilise un robot chien posé sur un tapis de jeu pour bebé et un algorithme appelé de « curiosité artificielle » qui mime le développement chez l'enfant en donnant une motivation au robot vers des interactions avec différentes parties de l'environnement de façon à maximiser son progrès dans sa capacité à prédire les conséquences de l'action. Il y a là une curiosité envers de l'information : lorsque quelque chose de non prévisible se produit, cela se présente comme de l'information que l'agent ne maîtrise pas et qu'il a donc besoin d'acquérir. Par la répétition, l'agent sera enfin capable de maîtriser cette information, donc de prédire ce qui va se produire (par exemple, toucher plusieurs fois un même objet pour reproduire un son). Un graphique de cette étude illustre la dynamique qui se produit au cours du temps chez ce genre de robot. En regardant le rapport entre le temps et l'erreur de prédiction pour quatre activités différentes, les résultats ont montré que sur une des activités (Activité 1), il y avait toujours une très forte erreur de prédiction. Normalement, la conséquence devrait être une négligence de la part du robot face à une tâche d'apprentissage trop difficile. À côté de cela, les activités 2 et 3 sont moyennement prévisibles et l'activité 4 très simple, donc trop prévisible. Entre deux activités moyennement imprévisibles (la 2 et la 3), l'idée est qu'il faudrait se focaliser d'abord sur celle qui permet de progresser plus rapidement, car c'est elle qui maximise la quantité d'informations acquises. Le robot va ainsi statistiquement consacrer beaucoup de temps à l'activité 3 et, dès que celle-ci commencera à être maîtrisée par le robot, elle deviendra de plus en plus prévisible. En conséquence, le robot va basculer vers l'activité 2 qui reste à découvrir. Un résultat intéressant de cette étude montre que ce genre de motivation intrinsèque chez le robot génère quelque chose qui ressemble à des séquences de comportements similaires à celles observées dans des phases développementales chez le bebé ou chez l'enfant.

Les risques liés à l'autonomie des robots

Ces exemples montrent qu'à partir du moment où l'on développe un certain niveau d'autonomie décisionnelle chez des robots, que ce soit dans le but d'une meilleure compréhension de la cognition humaine ou pour développer des applications, on s'expose au risque que le robot puisse agir de manière imprévisible et potentiellement dangereuse pour les humains. Dans une des nouvelles d'Isaac Asimov sur les robots, on trouve un robot appelé Tony, dont les connexions du cerveau sont décrites comme « très déterminées », contenant des éléments qui lui permettent de respecter les lois de la robotique et de manipuler la langue anglaise ainsi que "suffisamment d'autres notions pour accomplir le travail auquel il est destiné". Or, ce serait une erreur de qualifier ce genre de robot d'intelligent ou d'autonome : il s'agit tout au plus d'un bon robot industriel, capable de répéter des tâches pré-programmées avec des petites capacités d'adaptation. À l'inverse,

[8] Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive neuroscience*, 6(4), 187-214

[9] Oudeyer, P. Y. (2018). Computational theories of curiosity-driven learning. arXiv preprint arXiv:1802.10546.

dans d'autres nouvelles d'Asimov, on trouve le robot Lenny, qui a la taille d'un enfant et pour lequel le personnage du Dr. Susan Calvin, de l'U.S.Robots, éprouve un sentiment maternel. Ce robot apprend par lui-même et, en conséquence, fait des essais et des erreurs. Ceci constitue la clé de son apprentissage et de son autonomie : sans ces essais et erreurs, il ne pourrait pas acquérir de nouvelles informations, et donc pas progresser. Or, ces erreurs peuvent l'amener parfois à violer les lois de la robotique, comme par exemple à blesser involontairement un humain avant de connaître sa propre force physique, un peu comme cela peut arriver à un adolescent humain qui grandit trop vite et qui ne maîtrise pas son corps. La question éthique est de savoir si les torts et blessures mineurs, infligés aux humains qui l'entourent au cours de son apprentissage, sont tolérables et acceptables. Si l'on revient sur les lois de la robotique évoquées dans ces nouvelles, on peut se demander si elles seraient souhaitables aussi dans la société humaine, au-delà de la simple fiction. Pour les rappeler : selon la première loi, un robot ne peut porter atteinte à un être humain ni permettre qu'un être humain soit exposé à un danger en restant passif. La deuxième loi indique qu'un robot doit obéir aux ordres que lui donne un être humain sauf si de tels ordres entrent en conflit avec la première loi. La troisième loi affirme qu'un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.

Pour finir, concernant les conséquences sociales potentielles de l'autonomie décisionnelle chez les robots, il faut d'abord remarquer qu'il n'y a pas besoin de grandes capacités cognitives chez un agent artificiel, pour que ses décisions autonomes posent déjà de sérieux problèmes éthiques. Les applications de la robotique dans le domaine militaire sont un exemple de ce principe : on est loin d'avoir obtenu de grandes capacités cognitives pour ces robots, mais depuis déjà plus de dix ans (voir par exemple l'accident qui s'est produit en Afrique du Sud en 2007), ils peuvent parfois prendre des décisions autonomes comme celle de tirer sur des cibles. Et ceci a déjà conduit à des accidents, et parfois à mort d'homme. Leur emploi pose donc de grandes questions. Dans l'introduction au numéro spécial d'*Intellectica* sur Éthique et sciences cognitives[10] (op. cit.), j'avais écrit que « l'autonomie croissante des systèmes d'armement est devenue une problématique centrale débattue au niveau international dès 2013, avec notamment un rapport de l'ONU auquel a contribué Raja Chatila et d'autres sur le développement d'armes autonomes pouvant, une fois activées, poursuivre leurs cibles et passer à l'exécution sans intervention humaine ». Ce type de système robotique est développé dans l'idée qu'il pourra être plus efficace qu'un humain : le programme informatique pourra plus rapidement détecter et viser la cible, et la suivre avec davantage de précision qu'un humain. Pourtant, cela soulève un sérieux problème éthique de savoir s'il est acceptable qu'une machine autonome puisse avoir le pouvoir de vie ou de mort sur un être humain. C'est pourquoi, un ensemble de chercheurs ont lancé l'initiative « Stop Killer Robots » au niveau de l'ONU (Righetti et al., 2018)[11]. Une question presque systématique et centrale pour tout roboticien concerne donc les applications militaires potentielles de ses travaux sur l'autonomie des agents artificiels.

Un autre enjeu à mettre en évidence est le risque d'impact sur la société en termes d'emploi et de conditions de travail entraîné par le développement de robots autonomes, dotés de capacités de planification de l'action, de navigation ou d'interaction sociale. Plus précisément, par leur contribution à l'automatisation de la production industrielle, les robots conduisent à une disparition des métiers les

[10] Khamassi Mehdi, Chatila Raja & Mille Alain (Eds), Éthique et sciences cognitives, *Intellectica*, 70, (pp.7-39), DOI: n/a.

[11] Righetti, L., Pham, Q. C., Madhavan, R., & Chatila, R. (2018). Lethal autonomous weapon systems [ethical, legal, and societal issues]. *IEEE Robotics & Automation Magazine*, 25(1), 123-126.

moins qualifiés, tout en ouvrant la possibilité de créer de nouveaux métiers pour la conception et le maintien de ces robots. Même s'il peut se rassurer en considérant que les robots pourront aider à libérer les humains des tâches ingrates, répétitives, aliénantes et dangereuses, le chercheur ne doit pas éluder la question des potentielles conséquences sociétales de ces robots. Dans l'article de 2018 de Pham et collaborateurs[12], il est souligné que, malgré leur potentiel, les robots sont perçus par une part de l'opinion publique comme une menace pour l'emploi, notamment en conséquence du fait que le profit réalisé grâce à l'automatisation de la production ne semble pas s'accompagner systématiquement d'une augmentation de programmes éducatifs ou du déploiement de moyens financiers qui permettraient aux employés non qualifiés d'atteindre un niveau de qualification plus élevé. Les auteurs soulignent que les chercheurs ne doivent pas négliger les questions politiques et éthiques - qui possède les robots, comment est organisé le système économique actuel dans lequel les robots seront déployés - pour pouvoir envisager une contribution sociétale positive sur le long terme de la recherche en robotique.

Enfin, certains rapports[13] de chercheurs en IA et robotique mentionnent le risque d'exploitation d'agents artificiels autonomes par des entités qui seraient sans contrôle démocratique. Il est donc important de réfléchir, en amont, sur l'anticipation possible d'une période où l'IA forte ferait partie de notre société, avec la présence d'agents artificiels doués de la même autonomie que nous ou de capacités cognitives dépassant l'humain, voire même manipuler l'humain en fonction de leurs propres objectifs. Ce genre de réflexion peut être en partie mené au sein de TESaCo, en imaginant des scénarios concernant les technologies dites émergentes.

[12] Pham, Q. C., Madhavan, R., Righetti, L., Smart, W., & Chatila, R. (2018). The impact of robotics and automation on working conditions and employment. *IEEE Robotics & Automation Magazine*, 25(2), 126-128.

[13] Voir par exemple le rapport The malicious use of AI : Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228. ; http://arxiv.org/abs/1802.07228.

DISCUSSION

Daniel Andler

Imaginer déjà des scénarios liés à l'IA forte et se demander si les robots de demain s'aligneront sur nos valeurs ne me semble pas pertinent pour l'instant. Ne serait-il pas d'ailleurs difficile d'identifier des valeurs qui puissent représenter l'humanité entière ? Des problèmes éthiques actuels et bien plus urgents se posent, un exemple parmi beaucoup d'autres étant la condition des micro-travailleurs de l'IA dont il était question pendant l'intervention d'Antonio Casilli au colloque de lancement de TESaCo[14]. Réfléchir aux super robots de demain apparaît alors comme une façon de se détourner de ces questions urgentes.

Florian Forestier

Je voudrais revenir sur cette idée, souvent discutée, que les robots remplaceraient le travail humain. Ce qui a beaucoup mis ce thème en avant, c'est une étude fameuse menée à Oxford par Frey et Osborne[15] et qui a été très médiatisée. Cependant, celle-ci a été beaucoup remise en cause, à la fois dans sa méthodologie, dans la manière dont on avait analysé la substituabilité d'un emploi. Le risque de substitution comme tel - un chômage directement lié à la robotisation - paraît moins probable aujourd'hui. Au contraire, comme le souligne Antonio Casilli, l'interférence très forte par la mise en conformité de certaines tâches de travail même très qualifiées avec des séquences robotiques revient à une dégradation de la qualité du travail et s'apparente à la prolétarisation[16].

Mehdi Khamassi

Le fait de développer des technologies et de modifier la façon de travailler est en effet intrinsèque à l'humain. On peut même évoquer des visions plus optimistes comme celle des roboticiens qui considèrent que ce serait un bénéfice de soulager les humains des tâches aliénantes et répétitives par la robotisation de celles-ci. Certains auteurs, comme Bernard Stiegler, poussent encore plus loin cette réflexion en proposant des solutions comme celle d'automatiser le plus de tâches possibles de manière à ce que les humains puissent se consacrer à des activités intellectuelles ou artistiques, ou puissent tout simplement se consacrer aux projets et loisirs qu'ils affectionnent. Il est intéressant de prendre aussi en compte ces perspectives.

Serena Ciranna

Il s'agit peut-être d'une question marginale, mais je me demandais si le développement d'une IA forte, et notamment d'une autonomie décisionnelle chez les robots, pourrait en même temps accélérer la réalisation de certains scénarios transhumanistes, dont la possibilité d'un humain augmenté par l'IA ?

---

[14] Casilli A. «Qu'est-ce qu'une intelligence artificielle « réellement éthique » ?» *Les Cahiers de TESaCo* n°1.
[15] https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf?link=mktw

[16] Voir dans ce volume Florian Forestier, « Quel droit pour le management algorithmique ? » *Les Cahier de TESaCo* n°2

Mehdi Khamassi

En ce qui concerne l'augmentation de l'humain, ce serait plutôt une question liée à l'IA faible, notamment à l'extraction de connaissances de données pour assurer des performances humaines plus rapides. D'ailleurs, si cette technologie peut aider les humains à prendre de bonnes décisions en pleine transparence, il s'agit d'une bonne utilisation de l'IA. Par contre, se doter d'une IA forte pour remplacer la prise de décision chez l'homme reviendrait à laisser un agent extérieur prendre une décision à notre place, ce qui pose des problèmes pratiques et éthiques.

Daniel Andler

Dans le deuxième cas, il s'agirait, si l'on peut dire, d'un transhumanisme *fort*, au sens où la prothèse devient une greffe et l'individu n'est plus ce qu'il était avant. Dans ce dernier cas, la question de l'autonomie se pose vraiment.

Mehdi Khamassi

S'agissant par exemple des prothèses de mémoire, ce serait difficile d'imaginer quelles formes d'augmentation de l'humain peuvent impacter notre autonomie de décision. Je doute, en ce sens, que la partie de la robotique consacrée à développer une autonomie décisionnelle chez les robots puisse contribuer au projet transhumaniste.

Daniel Andler

Je voudrais soulever le problème de l'autonomie théorique et réelle qui se pose évidemment dans le cas de la prise de décisions médicales ou dans le domaine militaire. L'idée de garder toujours l'homme dans la boucle n'est en pratique pas toujours possible à réaliser dans des situations d'urgence. Il s'agit d'une question importante, d'autant plus qu'un biais d'automatisation existe, de sorte que l'opérateur humain a tendance à faire trop confiance à une solution qui ne lui est même pas imposée mais suggérée par un processus automatisé. L'autonomie formelle est alors respectée mais la tendance de l'agent humain à suivre aveuglément ces systèmes met en danger son autonomie.

Mehdi Khamassi

À ce sujet, on peut même ajouter que la façon dont la solution proposée est présentée et cadrée peut influencer la capacité d'acceptation de l'humain et par conséquent réduire son autonomie décisionnelle.

Margaux Berrettoni

La robotisation et l'automatisation du travail posent aussi des problèmes de cybersécurité. Si un robot est un ordinateur qui interagit avec son environnement, et si tout ordinateur est susceptible de se faire hacker, alors les robots sont eux aussi vulnérables. Que se passerait-il alors si un individu ou un groupe prenait le contrôle de robots autonomes ? Pour les systèmes informatiques, nous pouvons déjà évaluer l'ampleur des risques causés par un piratage. Par exemple, dans leur enquête sur

les nouvelles guerres[17], les journalistes Étienne Huver et Boris Razon sont revenus sur l'attaque du malware NotPetya qui avait frappé en juin 2017 une partie de l'Ukraine, dont la centrale nucléaire de Tchernobyl, des banques, des entreprises, des aéroports, des hôpitaux mais aussi d'autres entreprises européennes, comme Saint-Gobain en France et leur site des hauts fourneaux de Pont-à-Mousson en Lorraine. Toutes ces entreprises ont perdu le contrôle de leurs ordinateurs en l'espace de quelques heures et se sont retrouvées paralysées dans leurs activités. Cet exemple a d'abord montré une certaine fragilité des structures informatiques sur lesquelles reposent aujourd'hui une grande partie de nos activités. Dans le cas du site des hauts fourneaux en Lorraine, les installations industrielles qui produisent de l'acier en chauffant le minerai à 1600°C étaient gérées entièrement par des ordinateurs. Une fois ceux-ci paralysés, il a fallu s'appuyer sur l'ancien savoir-faire d'opérateurs humains qui connaissaient encore le processus manuel inusité depuis une trentaine d'années avec l'informatisation des processus. Ce sont ces actions qui ont permis d'éviter les explosions dans la centrale. En plus de la fragilité des systèmes informatiques, cet exemple pose aussi la question de la pérennité voire de la survie des savoir-faire.

Daniel Andler

Dans le magnifique livre de David Mindell sur les robots[18], on trouve des exemples concernant le pilotage des avions, notamment par rapport à l'accident du vol Rio-Paris. Il semble bien qu'il soit dû aux erreurs des jeunes pilotes peu expérimentés. Les Airbus sont particulièrement automatisés et ils ont réagi de travers face à une situation inattendue...

Mehdi Khamassi

Effectivement, cela pose le problème d'une automatisation partielle : une phase intermédiaire dans laquelle l'humain est encore dans la boucle et qui peut présenter des risques dus justement à cette interaction entre humain et machine. La question me semble aussi se poser dans le cadre du déploiement envisagé de véhicules autonomes : l'auto-régulation du travail ne peut alors fonctionner que si tous les véhicules répondent à la même logique pré-programmée ; soit si aucun véhicule présent n'est piloté par un humain. En d'autres termes, il se peut que dans certains cas l'optimum soit atteint seulement au prix d'une totale automatisation. Est-ce acceptable ?

[17] Étienne Huver, Boris Razon, Les nouvelles guerres. Sur la piste des hackers russes, Stock, 2019
[18] David A. Mindell, Our Robots, Ourselves: Robotics and the myths of autonomy, Viking, 2015