Giornale di filosofia, vol. 2/2021 • ISSN 1827-5834 • DOI 10.7413/1827-5834017

Martin Gibert[*]

Automatiser les théories morales

Introduction

Imaginons qu'on veuille construire un agent conversationnel – un *chatbot* – doté d'une intelligence artificielle ou boosté à l'intelligence artificielle. De nombreuses questions morales se posent. Son design conversationnel devrait-il être distinct selon que l'utilisateur est adulte ou enfant? Faut-il lui assigner un genre – en lui donnant une voix masculine, féminine ou neutre? Quelle finalité lui assigner? Maximiser le temps d'utilisation (en captant l'attention), améliorer la vie de l'utilisateur, améliorer la vie de tous?

Ces questions relèvent de l'éthique des machines ou de l'éthique des algorithmes dans la mesure où elles se posent lors de la programmation. Elles peuvent sans doute être informées par des considérations qui s'appuient sur des théories morales, mais elles ne requièrent pas qu'on automatise une théorie morale. Ces questions peuvent être résolues une fois pour toutes: en décidant par exemple de fabriquer un chatbot pour adulte à la voix neutre qui ne captera pas leur attention sans bonnes raisons.

D'autres questions, en revanche, ne peuvent être résolues de cette manière. Notre agent conversationnel devra-t-il dire la vérité à un mari soupçonneux et peu respectueux de la vie privée de sa conjointe? Dis Siri, ma femme était-elle seule à la maison aujourd'hui? Et que devra-t-il répondre à un jeune enfant qui veut savoir si le père Noël existe? Plus généralement, est-il acceptable qu'un chatbot mente à celui ou celle qui l'utilise? Un mensonge par omission est-il acceptable?

Ces questions présentent un degré supérieur de complexité parce qu'elles demandent que ce soit en quelque sorte le système d'IA qui, dans un contexte donné, prenne une décision: mentir ou non. Elles demandent qu'on en passe par une sorte d'automatisation de la morale. Les voitures autonomes dans des situations de type dilemme du tramway sont souvent citées pour montrer la nécessité d'automatiser la prise de décision

[*] Université de Montréal

morale[1]. Mais les besoins sont multiples, notamment pour les systèmes de recommandations. Voilà donc tout le défi: comment programmer un système d'intelligence artificielle (robots, chatbots, algorithmes) pour qu'il agisse moralement – ou à tout le moins plus moralement que ne le ferait ordinairement un être humain. On comprend alors tout l'intérêt qu'il pourrait y avoir à automatiser nos théories morales.

L'éthique des algorithmes et les agents moraux artificiels

C'est quand même tout un défi. Il s'agit de traduire les théories morales et les procédures de décision en code informatique, de les transcrire en algorithmes. On peut parler d'éthique des algorithmes pour qualifier ce sous-domaine de l'éthique appliquée à l'IA qui s'intéresse aux questions morales posées par la programmation d'un algorithme[2]. Est-il seulement possible de coder la morale? Comment des théories morales très générales comme le déontologisme, l'utilitarisme ou l'éthique de la vertu pourraient-elles être suffisamment formalisées pour recevoir une traduction algorithmique? On voit mal un robot parvenir à répliquer toute la subtilité d'une prise de décision morale.

Dans le domaine de la normativité, on peut cependant penser que l'opération n'est pas si inédite que cela. Après tout, le droit a bien quelque chose à voir avec les algorithmes. Le concept même d'algorithmes développée par le mathématicien persan Al-Khwârizmî (dont le nom latinisé est à l'origine du terme algorithme) serait initialement apparu pour aider les juges à résoudre facilement des casse-têtes juridiques liés notamment à des successions[3]. Les juges pouvaient ainsi suivre "comme des robots" les différentes étapes proposées par Al-Khwârizmî pour déterminer les parts d'un d'héritage.

Pourtant, quand bien même le droit essaye de formaliser le plus précisément possible les règles à suivre pour trancher les différends, le recours à des juges et à leur sens de la justice, le besoin d'une interprétation de l'esprit de la loi marquent bien l'impossibilité de s'en tenir à la lettre de la loi. Par comparaison, on devine l'ampleur de la tâche pour la philosophie morale qui, loin d'avoir le niveau de formalisation du droit, cherche à offrir des standards de conduite aux ingénieur.es chargé.es de programmer moralement des systèmes d'IA.

[1] S. Nyholm, *The ethics of crashes with self-driving cars: a roadmap*, in "I. Philosophy Compass" 13, 7, 2018.
[2] M. Gibert, *Faire la morale aux robots: une introduction à l'éthique des algorithmes,* Montréal Atelier 10, Flammarion, Paris 2021.
[3] Guerraoui R., Hoang N. L., *Turing à la plage*, Dunod, Malakoff 2020.

Devant une telle difficulté, la tentation du renoncement est grande. Puisque les choses sont si compliquées, pourquoi ne pas assumer d'emblée que seuls des humains sont aptes à prendre des décisions morales? Pourquoi ne pas admettre qu'il est vain, voire dangereux de demander à un système d'IA de prendre des décisions à notre place? L'éthique ou la morale (suivant la tradition analytique, je ne ferai pas de distinction entre ces deux notions) ne seraient tout simplement pas automatisables.

Il existe un autre argument pour un tel renoncement: c'est que les systèmes d'IA ne sont pas des agents moraux au sens fort. Ils ne peuvent être considérés comme responsables de leurs décisions. À quoi bon, par exemple, louer (ou blâmer) un agent conversationnel ou l'algorithme de YouTube pour leurs bonnes (ou mauvaises) décisions? Contrairement à des agents moraux humains, ni l'un ni l'autre ne sont des entités conscientes et il semblerait absurde de leur imputer une quelconque responsabilité morale – alors qu'il semblerait parfaitement légitime de le faire pour celles et ceux qui ont programmé ces systèmes d'IA.

Comme le remarque par ailleurs Rodogno et Nørskov, ces robots ne maîtrisant pas le langage des raisons, ils ne pourront pas se justifier pour leurs actions et une voiture autonome ne pourra pas s'excuser auprès des parents des victimes d'un dilemme[4]. On peut regretter cette perte d'un aspect central de nos pratiques morales. On peut penser qu'il existe une irréductibilité du langage des mathématiques, qui est au cœur des algorithmes, à celui des raisons, qui est au cœur de l'éthique. Mais il n'en demeure pas moins que nous avons besoin d'une éthique des algorithmes.

Renoncer à traduire l'éthique en code informatique, renoncer à automatiser les théories morales ne résoudra évidemment pas le problème. En effet, les chatbots ou l'algorithme de YouTube, dans la mesure où ils sont au moins en partie autonomes, peuvent être qualifiés d'agent moraux artificiels (AMA), selon l'expression de Wallach et Allen qui présentent ainsi le concept:

> Aujourd'hui, les systèmes [automatiques] s'approchent d'un niveau de complexité qui, selon nous, exige qu'ils prennent eux-mêmes des décisions morales, qu'ils soient programmés avec des "sous-routines éthiques" pour reprendre une expression de Star Trek. Cela va élargir le cercle des agents moraux au-delà des humains à des systèmes artificiellement intelligents, que nous appellerons des agents moraux artificiels[5].

[4] R. Rodogno, M. Nørskov, *The automation of ethics: the case of self driving cars*, (a cura di) C. Hasse and D. M. Søndergaard, in "Designing Robots – Designing Humans", Routledge, Londres (à paraitre).

[5] W. Wallach, C. Allen, *Moral machines: teaching robots right from wrong*. Oxford University Press, Oxford 2008.

C'est parce que les agents moraux artificiels doivent prendre des décisions – et quand bien même ils ne sont pas imputables de ces décisions – qu'une programmation morale s'impose et qu'il parait judicieux de se demander si les théories morales traditionnelles peuvent fournir une architecture.

Un cas paradigmatique d'AMA serait une voiture autonome qui, face à un accident inévitable – à la manière d'un dilemme du tramway – aurait à "choisir" entre tourner à gauche ou à droite mettant ainsi en danger des personnes distinctes. Or, quand bien même ce robot de transport n'est pas un agent moral au sens fort, nul doute que ses prises de décisions auront des conséquences moralement significatives. La voiture autonome fera ce qu'on lui demande de faire et il est parfaitement légitime que les ingénieur.es et les philosophes collaborent pour réaliser de telles AMAs.

Il faut savoir que les trois théories morales proposent des réponses distinctes à la question de base de l'éthique, soit "quelle est la bonne chose à faire?". On peut ainsi dire que le déontologisme va traiter la question en se focalisant sur l'*action*, l'utilitarisme sur les *conséquences* de l'action et l'éthique de la vertu sur l'*agent*. Ces perspectives peuvent évidemment converger pour évaluer une action donnée. Ainsi, un déontologiste kantien pourra condamner le mensonge d'un gradé de l'armée française dans l'affaire Dreyfus parce qu'il est tout simplement mal de mentir, un utilitariste parce que cela nuit au bien-être de Dreyfus, décrédibilise l'armée française et renforce l'antisémitisme et un éthicien de la vertu parce qu'un agent vertueux devrait être honnête.

Mais ces trois théories sont d'autant plus intéressantes pour notre analyse qu'il leur arrive de diverger et de plaider pour des actions distinctes comme les manuels d'éthique ne manquent pas de le rappeler (par exemple lorsque l'action de mentir ou la malhonnêteté d'un agent ont de bonnes conséquences). Peut-on parvenir à formaliser ces divergences? Cela peut-il alors faire du sens que de parler d'agents moraux artificiels (ou de robots) à l'architecture déontologiste, utilitariste ou arétaïque, c'est-à-dire relative à l'éthique de la vertu?

Dans les trois prochaines sections, je vais soutenir qu'il existe effectivement des manières d'automatiser, à différents degrés, le déontologisme, l'utilitarisme et l'éthique de la vertu.

Automatiser le déontologisme

On peut penser que le déontologisme est la plus automatisable des théories morales. Il existe en effet une affinité spontanée entre cette théorie normative et la notion d'algorithmes. On la présente souvent comme

une approche centrée sur des règles: une action n'est pas correcte moralement si elle enfreint une règle (comme l'interdit du meurtre ou du mensonge) et correcte dans le cas contraire. Or, l'algorithme d'un système d'intelligence artificiel peut tout à fait intégrer de telles règles.

Susan et Michael Anderson (elle est philosophe et son mari ingénieur) peuvent être considérés comme des pionniers. Dans un article du *Scientific American* publié en 2010, ils expliquent comment ils ont implanté des règles morales à NAO, une petite machine humanoïde utilisée comme robot de soin (*carebot*)[6]. Quand bien même les possibilités de NAO sont limitées, on peut lui demander d'apporter ses médicaments à un patient aux intervalles de temps prescrits par le médecin.

NAO étant équipé d'une caméra, il est capable de savoir qu'un patient n'a pas pris ses médicaments. La sous-routine éthique – pour reprendre l'expression de Wallach et Allen – qui intéresse les époux Anderson c'est de déterminer quand NAO doit insister auprès d'un patient. Bien qu'il ne s'agisse pas à proprement parler d'un dilemme, la situation est moralement délicate: les patients sont des adultes responsables et l'insistance du robot – tout comme celle des membres du personnel médical– pourrait relever d'un paternalisme inapproprié. Comment programmer NAO pour qu'il soit un *bon* robot de soin?

Les époux Anderson se sont inspirés d'un déontologiste (pluraliste) fameux, le britannique David Ross[7]. Selon sa théorie des devoirs *prima facie* il existe des règles morales de base à respecter comme être honnête, tenir ses promesses ou ne pas faire de tort à autrui. La prise de décision morale consiste alors à combiner les différents devoirs – Ross utilise la métaphore d'une somme vectorielle – afin de déterminer l'action à accomplir dans chaque situation.

Un *bon* robot de soin, selon Susan et Michael Anderson, devrait ainsi respecter les devoirs suivants, par ordre de priorité: 1) la non malfaisance: ne pas mettre la vie des patients en danger, 2) le respect de l'autonomie: ne pas aller à l'encontre de la volonté des patients, 3) la bienfaisance: aider les patients à prendre leurs médicaments. Cela signifie que NAO devrait éviter d'être paternaliste à moins que la vie des patients ne soit en danger. Il s'ensuit un algorithme relativement simple qui détermine la marche à suivre pour NAO en suivant une suite de conditionnelles du type "si A alors B" et "si non-A alors C".

Mais pour être complet, l'algorithme du robot devrait aussi tenir compte des informations fournies par le médecin: à quel moment prendre les médicaments? Que risquent les patients s'ils ne les prennent

[6] M. Anderson, S.L. Anderson, *Robot be good*, in "Scientific American", 303, 4, October 2010, pp. 72-77.
[7] W.D. Ross, *The right and the good*. Oxford: Clarendon Press, Oxford 1930.

pas? Combien de temps peuvent-ils récidiver avant que la situation ne s'aggrave? On le voit, programmer des règles va souvent de concert avec fixer des seuils. Il pourrait par exemple être acceptable que le robot insiste une fois sur quatre auprès des patients qui oublient de prendre leurs pilules, mais pas davantage.

De façon générale, traduire le déontologisme en algorithmes, c'est apprendre à un AMA à subsumer un cas particulier ("X est ou n'est pas un mensonge") sous une règle normative ("il ne faut pas mentir"). Une telle approche soulève plusieurs questions.

D'abord quelles normes choisir? En supposant qu'un chatbot puisse tromper sciemment un utilisateur, faut-il vraiment le lui interdire? Cela inclut-il les mensonges par omission? Ne faudrait-il pas plutôt programmer le chatbot à mentir si cela peut sauver des vies, comme dans l'exemple célèbre de Kant qui attira les foudres de Benjamin Constant? En définitive, la question du choix des normes concerne bien sûr tout l'édifice de l'éthique et renvoie à des enjeux en épistémologie morale: comment reconnaître les normes morales à implémenter.

À cet égard, on voit combien il serait difficile de construire un robot déontologiste universel, c'est-à-dire conçu pour pallier tout type de décision, dans toutes les situations moralement "à risque". Cela supposerait une carte morale extrêmement complexe de toutes les règles morales ou, pour reprendre la métaphore de Ross, cela demanderait de connaître l'ensemble des vecteurs moraux qui doivent guider nos actions à chaque instant. Inutile de dire que nous sommes loin du compte et qu'il est peu probable qu'un consensus émerge au sein des philosophes moraux (même en se restreignant aux déontologistes) pour une liste précise de normes à implémenter. C'est évidemment là une première difficulté de taille pour la mise en algorithme d'une approche déontologiste.

Supposons toutefois que l'on s'accorde sur quelques normes comme "ne pas tuer" ou "respecter les droits fondamentaux des individus". Plus généralement – et si on laisse de côté les zones où l'éthique diverge du droit – on peut penser qu'un robot ou un algorithme déontologiste devrait respecter les codes juridiques.

S'il n'existait qu'une seule règle à respecter, si l'éthique était mono-normative, la mise en algorithme du déontologisme serait relativement aisée. Malheureusement, comme le suggèrent les dix commandements, les sept devoirs *prima facie* de Ross[8], les quatre principes de la bioéthique ou les trois règles que les Anderson ont implémentées dans NAO, il y a habituellement *plusieurs règles* à combiner.

Telle est la seconde difficulté de la mise en algorithme du déontolo-

[8] *Ibid.*

gisme: comment combiner une multiplicité de règles, c'est-à-dire comment les hiérarchiser ou les pondérer les unes par rapport aux autres? Traduit dans la métaphore de Ross, cela signifie qu'il ne nous faut pas seulement connaitre quelles sont les normes morales en jeu, mais aussi quelles sont leurs *valeurs* et leurs *directions* relatives puisqu'on entend en faire la somme vectorielle.

Une troisième difficulté est inhérente au caractère généraliste des règles confrontées à la singularité des situations que rencontrent les robots. Cela correspond aux critiques que le particularisme moral – une théorie métaéthique – adresse au généralisme moral. Comme l'écrit Jonathan Dancy, selon le particularisme moral l'arbitrage parfaitement moral "aurait besoin de bien plus que de la maîtrise d'une gamme appropriée de principes et de la capacité de les appliquer[9]". Pourquoi? Parce que la moralité est irréductible à des règles générales, parce que le niveau de granularité des raisons morales est bien plus fin que les gros sabots du généralisme.

Dans cette perspective on ne peut affirmer qu'*il est mal de tuer*, car la peine de mort n'est pas forcément immorale ou parce qu'on peut tuer en cas de légitime défense. On peut, tout au plus, dire qu'*il est généralement mal de tuer*, mais on doit éviter de penser la prise de décision morale en termes de principes (absolus ou *pro tanto*), même hiérarchisés. Bref, pour le particularisme moral, il y a toujours des exceptions aux principes et des exceptions aux exceptions. Pour filer la métaphore rossienne, on peut dire que selon les particularistes, la valeur et la direction des normes ne sont pas fixes: elles changent selon la situation. Dans ces conditions, on comprend que la mise en algorithme de l'éthique – à tout le moins selon une approche déontologiste – paraît aussi vaine qu'erronée.

En fait, le débat entre particularistes et généralistes (ou *principlistes*) moraux traverse toutes les théories normatives puisque le particularisme remet en question l'idée même de règle ou de principes moraux (mais non pas qu'il existe des vérités morales).[10] De ce point de vue, les critiques particularistes pourraient aussi bien s'adresser à l'utilitarisme qu'à l'éthique de la vertu. Mais puisque le déontologisme met la notion de règle au centre de sa formalisation de la morale, on comprend qu'il se retrouve sur le front principal.

En résumé, l'automatisation du déontologisme doit affronter un problème d'identification et de pondération des normes morales, mais aussi la critique particulariste de la prise de décision morale par des règles/

[9] J. Dancy, *Moral Particularism*, in "The Stanford Encyclopedia of Philosophy", a cura di Edward N. Zalta, 2017.
[10] M. Ridge, S. McKeever, *Moral Particularism and Moral Generalism*, in "The Stanford Encyclopedia of Philosophy", (a cura di) Edward N. Zalta, 2016.

principes. Dans les faits, il n'en est pas moins possible d'implémenter des règles et des étapes à suivre, bref un algorithme, dans des AMAs. La programmation du robot NAO par les époux Anderson en atteste et, bien que leur ambition soit modeste, on peut y voir une preuve de concept de robot déontologiste.

La recherche en éthique des machines et des algorithmes qui se poursuit dans cette veine[11] peut se prévaloir d'une maintenant relativement longue histoire des "systèmes experts", ces logiciels cherchant à automatiser les compétences d'experts humains, de la calculatrice au jeu d'échecs électronique en passant par les modules des pilotes automatiques dans l'aérospatial. Avec ses hiérarchies structurées de règles et son code "écrit à la main", la technique des systèmes experts semble être taillée sur mesure pour le déontologisme – et la mieux parée pour son automatisation.

En pratique, l'automatisation du déontologisme semble surtout fonctionner dans un environnement contrôlé où l'on peut réduire la prise de décision morale à l'application de quelques règles bien définies.

Automatiser l'utilitarisme

Il existe une bonne manière de répondre à la critique particulariste à l'endroit de l'approche déontologique. C'est de sortir de la logique des principes ou des règles qui se combinent pour conclure qu'il faut faire ceci ou cela (les raisons *pro tanto* ou les vecteurs qui s'additionnent). Sortir de la logique d'une "programmation à la main" des règles et de leurs exceptions. Bref, sortir des systèmes experts: c'est ce à quoi l'on va assister avec l'automatisation de l'utilitarisme.

Avant toute chose, il faut expliciter le clivage moral essentiel entre le déontologisme et l'utilitarisme, cette théorie morale développée par Jeremy Bentham et qui appartient à la famille conséquentialiste. L'enjeu principal est de savoir si l'on doit respecter une règle, *quelles que soient ses conséquences* (en termes de bien être) ou si l'on peut mentir, briser une promesse ou tuer une personne *pour autant que cela entraîne de bonnes conséquences*. On pourrait y voir le clivage entre une logique binaire, celle du déontologisme – puisqu'une action respecte *oui ou non* une norme morale – et une logique du *plus ou moins* puisque l'utilitarisme qui vise la réalisation d'un objectif, à savoir la maximisation du bien-être.

Programmer un algorithme en visant un objectif s'avère très différent de le faire en combinant une multiplicité de règles. Cela permet notamment d'utiliser une technique d'IA qu'on nomme l'apprentissage auto-

[11] Voir par exemple: R. Tappl (a cura di), *A Construction Manual for Robots' Ethical Systems*, Springer, New York 2015.

matique. Dans ce cas, aucune règle ne préexiste (ce qui devrait tenir les particularistes tranquilles) mais la procédure de décision relève plutôt d'un apprentissage. Avec l'apprentissage profond en particulier, qui s'appuie sur des milliers de connexions dans des réseaux de neurones artificiels, la notion de règles disparaît mais l'algorithme n'en parvient pas moins à l'objectif qu'on lui a assigné.

Passer d'une approche en termes de règles à une approche en termes d'objectif facilite également la mobilisation d'une pensée bayésienne. On peut compléter un objectif avec un certain degré, et on peut être plus ou moins certains qu'on y parviendra. Apprendre pour un algorithme comme pour un cerveau biologique, c'est devenir une bonne machine à prédire.

On peut lire les plans très aboutis d'une automatisation de l'utilitarisme dans le fascinant livre de Stuart Russell, *Human Compatible* – en particulier dans les chapitres 7 et 9. Le célèbre chercheur en IA commence par assigner un objectif général aux systèmes d'IA: "Les machines sont bénéfiques dans la mesure où *leurs* actions sont censées atteindre *nos* objectifs[12]". Quant à l'intelligence d'une entité, c'est en gros sa capacité à atteindre ce qu'elle veut, étant donnée les informations qu'elle possède, par exemple grâce à sa perception du monde. C'est sa capacité à agir avec succès, notamment en termes évolutionnaires, mais pas seulement. En ce sens très inclusif peuvent être dites intelligentes des entités aussi diverses que les virus, les plantes, les animaux, les robots et les normaliens.

Russell note d'ailleurs la ressemblance entre le mécanisme à l'œuvre dans les cerveaux que constitue le circuit de la récompense et l'apprentissage par renforcement utilisé en IA. Avec cette dernière technique d'IA, il s'agit "d'apprendre à partir de l'expérience directe des signaux de récompense dans l'environnement, tout comme un bébé apprend à se tenir debout à partir de la satisfaction de se tenir droit et de la sanction de tomber[13]". C'est ce type de technique qui a été utilisé pour construire Alphago, un logiciel devenu champion de go.

Dans la seconde partie de son livre, Russell développe une architecture pour les AMA qui se présente explicitement[14] comme une

[12] S.J. Russell, *Human Compatible: AI and the Problem of Control*, Allen Lane, Londres 2019, p. 23.
[13] Ivi, p. 66.
[14] Il exclut explicitement les autres théories morales: "En l'absence de toute preuve de conscience de soi de la part des machines, je pense qu'il est peu judicieux de construire des machines qui sont vertueuses ou qui choisissent des actions conformes aux règles morales si les conséquences sont hautement indésirables pour l'humanité. En d'autres termes, nous construisons des machines pour provoquer des conséquences, et nous devrions

tentative de programmation de l'utilitarisme des préférences (dans la version de John Harsanyi[15]). L'IA peut nous aider à identifier les préférences des gens – notamment grâce à l'apprentissage par renforcement inversé (*inverse RL*). Selon cette nouvelle méthode, explique Russell, "le seul objectif du robot est de satisfaire les préférences humaines, il ne les connaît pas initialement et il peut en savoir davantage en observant les comportements humains[16]". Et le chercheur semble optimiste:

> En principe, la machine peut apprendre des milliards de modèles prédictifs de préférence différents, un pour chacun des milliards de personnes sur Terre. Ce n'est vraiment pas trop demander aux systèmes d'IA du futur, étant donné que les systèmes de Facebook actuels gèrent déjà plus de deux milliards de profils individuels.[17]

L'étape suivante pour construire une "machine intelligente bénéfique" consiste à agréger les préférences de diverses personnes, de manière à avoir un agent (artificiel) impartial: "Un agent agissant au nom d'une population d'individus doit maximiser une combinaison linéaire pondérée de l'utilité de ces individus[18]". Et lorsque des préférences sont en conflit, c'est l'algorithme qui tranche pour maximiser le bien-être de façon impartiale. On notera enfin que Russell insiste beaucoup sur l'importance des probabilités dans l'architecture des AMAs. Pour être efficaces, un système d'IA doit être un agent bayésien, qui ajuste et réajuste en permanence le modèle qu'il se fait de son environnement en termes de prévisibilité et d'incertitude.

Évidemment, l'automatisation de l'utilitarisme des préférences n'exonère en rien cette théorie des critiques qui lui sont habituellement adressées (sans parler des critiques plus générales adressées à l'utilitarisme comme son indifférence aux droits humains). Nos préférences peuvent évoluer, être irrationnelles, à court terme, purement dépendantes du contexte, voire sadiques. Russell, qui proposent un certain nombre de solutions pour filtrer les préférences irrationnelles, reconnaît certains points d'achoppement: "par exemple, les machines pourront

préférer construire des machines qui provoquent des conséquences que nous préférons. Cela ne veut pas dire que les règles morales et les vertus ne sont pas pertinentes; c'est juste que, pour l'utilitariste, elles sont justifiées en termes de conséquences et de la réalisation plus pratique de ces conséquences". Ivi, p. 234.

[15] J. Harsanyi, *Morality and the Theory of Rational Behavior*, in "Social Research: An International Quarterly 44", 4, 1977, pp. 623-656.

[16] Russell, *op. cit.*, p. 230.

[17] Ivi, p. 194.

[18] Ivi, p. 238.

avoir à traiter différemment celles et ceux qui préfèrent activement que les autres souffrent."[19]

D'autres chercheurs s'inspirent de l'utilitarisme pour développer des AMAs. Hoang et Mhamdi proposent une architecture modulaire très facilement adaptable à l'utilitarisme[20], tandis que Leben préconise un principe Maximin qui s'inscrit dans une logique conséquentialiste[21]. L'enthousiasme initial pour les systèmes experts et l'architecture déontologiste semble s'être aujourd'hui déporté vers l'apprentissage par renforcement et une architecture utilitariste. Avec son injonction précise et théoriquement mesurable – maximiser le bien-être – l'utilitarisme a tout pour plaire à celles et ceux qui, dans une optique conséquentialiste, voient la moralité comme la réalisation d'un objectif, lequel peut se traduire directement en algorithme sous la forme d'une *fonction d'objectif*.

Comme on l'a vu avec la question des préférences sadiques, l'architecture utilitariste doit affronter de sérieux défis. De plus, on pourrait la critiquer pour son esprit général: c'est une approche très "quantitative" qui semble mener à une disparition de l'humain dans le processus de décision.

Ce n'est assurément pas un reproche qu'on peut faire à la troisième approche. On va le voir, l'architecture arétaïque n'a de cesse de ramener l'humain au cœur de la conception des agents moraux artificiels.

Automatiser l'éthique de la vertu

Il y a peut-être une manière de résoudre le problème des préférences inappropriées auquel est confronté Russell. C'est de sélectionner une base d'apprentissage qui comporte de 'meilleures' préférences. On pourrait, par exemple, décider de ne pas tenir compte des préférences des psychopathes ou donner plus de poids aux préférences des personnes vertueuses. C'est exactement le type de stratégie qui prévaut avec l'architecture arétaïque.

En intelligence artificielle, l'apprentissage supervisé est, comme l'apprentissage par renforcement, une forme d'apprentissage automatique, qui se distingue donc des algorithmes "écrits à la main" des systèmes experts. Sa principale caractéristique, c'est de s'appuyer sur des exemples: de façon analogue à ce qui se passe dans un cerveau biologique, des neurones artificiels (des nœuds dans une structure mathématiques des algo-

[19] Ivi, p. 194.
[20] L.N. Hoang, *et al.*, *Le fabuleux chantier: Rendre l'intelligence artificielle robustement bénéfique*, EDP Sciences, Les Ulis 2019.
[21] D. Leben, *Ethics for robots: how to design a moral algorithm*. Routledge/Taylor & Francis, New York/London 2019.

rithmes) sont capables d'inférer – par une sorte d'induction automatique – des patterns dans les données d'apprentissage qu'on leur soumet.

C'est ainsi que fonctionne par exemple la reconnaissance d'image et de son. Alors qu'un enfant humain sera probablement capable de reconnaître une feuille d'érable ou un orignal après avoir vu deux ou trois photos, les meilleures techniques actuelles en apprentissage supervisé requièrent des milliers d'exemples de feuilles d'érable et d'orignaux, sinon des millions. Mais l'algorithme finit par reconnaître – avec un taux d'erreur qui peut être inférieur à celui d'un humain – les orignaux présents sur des photos. Voilà, en passant, pourquoi on parle tant d'IA depuis 2012: on s'est aperçu que l'apprentissage supervisé, longtemps marginalisé, tient ses promesses lorsqu'il est couplé avec des données massives. Ça fonctionne.

Dès lors qu'on s'intéresse à l'automatisation de la morale, il est difficile de ne pas voir le parallèle avec une théorie qui place l'exemplarité au cœur de la prise de décision morale. En effet, pour l'éthique de la vertu, non seulement nous devrions chercher à devenir de meilleures personnes en suivant l'exemple des modèles qui nous entourent, mais identifier la bonne action suppose de recourir à des exemples de personnes vertueuses. La philosophe néozélandaise Rosalind Hursthouse définit d'ailleurs ainsi l'action juste: "Une action est correcte [*right*] si elle correspond à ce qu'un agent vertueux ferait (en agissant selon son caractère) dans ces mêmes circonstances"[22].

Autrement dit, l'éthique de la vertu correspond à ce que la philosophe américaine Linda Zagzebski nomme une théorie *exemplariste,* c'est-à-dire qui fonde la moralité sur des exemples qui instancient le bien[23]. Pour Zagzebski, les théories morales peuvent être distinguées en fonction de l'attention ou du poids que chacune accorde à ces trois concepts moraux fondamentaux: le bien [*goodness*], privilégié par l'utilitarisme, le correct [*rightness*], privilégié par le déontologisme, et la vertu, privilégiée par qui vous savez. Si l'éthique de la vertu est plus volontiers exemplariste que les autres théories morales, c'est parce qu'elle place l'exemplarité des personnes vertueuses au cœur de son dispositif.

Mais comment détecter ces exemples moraux? Qui faut-il imiter, de qui faut-il s'inspirer lorsqu'on veut suivre l'éthique de la vertu? Ces questions valent tout autant pour un humain que pour un robot. Une idée consiste à s'appuyer sur la tradition et à analyser (de façon critique) ce qu'on dit des "saints moraux", ces personnes particulièrement ver-

[22] R. Hursthouse, *On virtue ethics*, Oxford University Press, Oxford 1999. La formule est celle d'Aristote.
[23] L. Zagzebski, *Exemplarist virtue theory*, in "Metaphilosophy", 41, 1-2, 2010, pp. 41–57.

tueuses, c'est-à-dire courageuses, honnêtes, bienveillantes etc. dont la vie et les actions – comme celles de Jésus ou de Bouddha – semblent dignes d'être imitées. Zagzebski suggère quant à elle de prendre pour guide une émotion: l'admiration morale.

Le chercheur sud-africain Bongani Andy Mabaso s'intéresse à la traduction en algorithme de cette approche exemplariste-arétaïque: "D'un point de vue purement computationnel, écrit-il, le problème de l'apprentissage à partir d'un exemple est plus facilement soluble que la programmation de la compréhension de concepts abstraits dans l'AMA[24]".

Plus concrètement, Mabaso considère un robot enseignant qui aurait à gérer une situation moralement délicate durant un cours. Supposons qu'une élève se mette à perturber la classe: comment le robot devrait-il réagir? Comme un enseignant moralement admirable répond l'éthique de la vertu! L'architecture arétaïque implique donc de recourir à une base de données contenant des exemples – comme des enregistrements vidéo – de situation de classe "étiquetés", c'est-à-dire avec un indicateur numérique qui évalue la manière dont l'enseignant a géré la situation.

Un avantage de cette approche, explique Mabaso, c'est qu'en plus d'être moralement acceptable – du moins pour les éthiciens de la vertu – elle peut répondre aux attentes spécifiques d'une communauté. Ainsi, dans le cas du robot-enseignant, on peut prendre les exemples moraux – les bons enseignants – au sein même de la communauté, qu'il s'agisse du district scolaire, de la nation ou d'autre chose.

Plus généralement, on peut penser que le robot vertueux devrait davantage inspirer confiance que ses concurrents – c'est son côté "plus humain". Parce qu'il se fonde sur l'exemplarité, qu'il se comporte comme le feraient des personnes moralement admirées, son intégration sociale sera d'autant facilitée: qui ne voudrait pas être entouré de robots honnêtes, avisés et bienveillants? Et cela vaut évidemment en tout premier lieu pour les robots sociaux.

Comme ses rivales déontologiste et utilitariste, l'architecture des AMAs arétaïque fait face à des défis de taille. Outre la difficulté déjà mentionnée d'identifier les personnes vertueuses, il y a celle de collecter des données moralement pertinentes et, enfin, celle de développer un algorithme de prise de décision qui corresponde à ce que ferait effectivement une personne vertueuse. Comment passer d'une multitude de données sur un grand nombre de personnes vertueuses à une procédure de décision qui respecte la diversité et la pluralité des exemples?

J'ai proposé ailleurs une architecture arétaïque pour une voiture au-

[24] B.A. Mabaso, *Artificial Moral Agents Within an Ethos of AI4SG*, in "Philosophy and Technology", 2020.

tonome faisant face à un dilemme de type tramway[25]. Comment programmer un tel robot de transport qui, face à un accident inévitable, devrait par exemple choisir entre écraser un enfant ou un vieillard? Une approche déontologiste – comme celle proposée par un comité allemand des transports en 2017 – impliquerait probablement un tirage au sort, considérant que l'enfant et le vieillard sont égaux, qu'ils ont les mêmes droits et que l'âgisme est une discrimination injuste. L'approche utilitariste, quant à elle, donnerait plus de poids à la vie de l'enfant dans la mesure où il va certainement vivre plus longtemps que le vieillard et "produire" ainsi plus de bien être (je simplifie beaucoup, on pourrait raffiner l'analyse utilitariste). Sauver l'enfant: c'est d'ailleurs l'intuition morale qui vient tête des personnes sondées dans de nombreuses régions du monde, comme l'a montré une vaste étude de psychologie morale[26].

Comment programmer un robot vertueux? Supposons qu'on possède des données précises sur les préférences morales de nos exemples de vertu (on leur a fait passer des tests avec plein d'expériences de pensée). Supposons que 50% d'entre elles estiment que la voiture autonome devrait écraser le vieillard, 30% qu'il devrait tirer au sort et 20% qu'il devrait écraser l'enfant. Si d'aventure une voiture était confrontée à un tel dilemme, je crois qu'une bonne manière de rendre compte de cette diversité d'intuitions serait de reproduire ces proportions dans l'algorithme de prise de décision. Ainsi, dans 50% des cas, la voiture sauverait l'enfant, dans 30% elle tirerait au sort et dans 20% des cas, elle sauverait le vieillard. On notera que ce type de programmation recourt au hasard, ce qui rend la décision imprévisible, mais tout en préservant une forme d'équité en raison des options proportionnelles.

On le voit, l'éthique de la vertu, avec toute sa dimension exemplariste, peut offrir à son tour une architecture morale aux agents moraux artificiels. Elle devra bien évidemment affronter des défis, comme la constitution d'une collection de personnes vertueuses ou, à défaut, une collection d'actions moralement exemplaires. Qui sait si notre chatbot ne pourrait pas devenir un meilleur compagnon conversationnel s'il "apprend à parler" à partir de texte ou de voix provenant de personnes plus vertueuses que la moyenne?

Notons enfin que pour "humaine" qu'elle soit, la programmation arétaïque ne suppose pas que les robots éprouvent des émotions. Ce sont les personnes vertueuses, les exemplaires de moralité, qui sont des êtres sensibles. Leurs émotions peuvent sans doute les aider à percevoir les diverses dimensions morales d'une situation (elles ont un rôle épistémique) et à prendre des décisions plus avisées. Le robot, lui, se contente d'imiter; il n'a pas besoin d'avoir mal pour savoir qu'il ne

[25] M. Gibert, *op. cit*.
[26] E. Awad, *et al.*, *The moral machine experiment*, in "Nature", 563 (7729), 59, 2018.

faut pas faire mal. Même dépourvu d'émotions, il peut tout à fait se comporter comme le ferait un agent sensible – pour autant qu'on le nourrisse avec des données appropriées. C'est pourquoi l'automatisation de l'éthique de la vertu passe par une meilleure compréhension de la psychologie des personnes vertueuses.

Conclusion

Comme je l'ai dit plus haut, on peut refuser par principe l'automatisation de la prise de décision morale. On plaidera que cette tâche ne devrait jamais être déléguée à des machines – et que la moralité est irréductiblement et essentiellement humaine. Mais ça n'aidera pas beaucoup celles et ceux qui doivent programmer un agent conversationnel à réagir de façon appropriée dans une situation moralement délicate – comme une question sur l'existence du Père Noël.

En définitive, on peut se demander si les philosophes et les ingénieur. es peuvent trouver une langue commune pour se parler. À cet égard, le genre d'affinité qu'on a pu constater entre les trois théories morales et trois techniques d'IA est plutôt un signe d'espoir. Certes, les ingénieures sont rarement des experts en éthique, et les philosophes moraux n'ont pas tellement l'habitude de réfléchir au niveau de détail et de rigueur des algorithmes. N'oublions pas non plus que la question de l'automatisation des théories morales, il y a quelques années encore, relevait des spéculations de la science-fiction. Mais le temps de s'y mettre semble bel et bien venu.

En fait, comme j'espère l'avoir montré dans cet article en présentant les travaux des époux Anderson avec le déontologisme, celles de Stuart Russell avec l'utilitarisme et celles de Bongani Andy Mabaso avec l'éthique de la vertu, la recherche sur l'automatisation des théories morales est déjà lancée. Quant à savoir laquelle de ces trois architectures morales prévaudra pour nos futurs agents moraux artificiels, à l'heure qu'il est, rien ne permet encore de le dire. Mais tout laisse à penser qu'une subtile combinaison des trois sera nécessaire.

Bibliographie

Awad E., *et al.*, (2018) *The moral machine experiment*, in "Nature", 563, 2018, pp. 59-64.

Anderson M., Anderson, S.L., *Robot be good*, in "Scientific American", 303, 4, octobre 2010, pp. 72-77.

Gibert M., *Faire la morale aux robots*, Montréal Atelier 10, Flammarion, Paris

2021.
Guerraoui R., Hoang N. L., *Turing à la plage*, Dunod, Malakoff 2020.
Harsanyi J., *Morality and the theory of rational behavior*, in "Social Research: An International Quarterly", 4, 1977.
Hursthouse R., *On virtue ethics*, Oxford University Press, Oxford 1999.
Leben D., *Ethics for robots: how to design a moral algorithm*, Routledge/Taylor & Francis, New York/London 2019.
Mabaso B.A., *Artificial Moral Agents Within an Ethos of AI4SG*, in "Philosophy and Technology", 2020.
Nguyen Hoang L., El Mhamdi E.M., *Le fabuleux chantier: rendre l'intelligence artificielle robusement bénéfique*, EDP Sciences, Les Ulis 2019.
Nyholm S., *The ethics of crashes with self-driving cars: a roadmap*., in "I. Philosophy Compass", 13, 7, 2018.
Ridge M., McKeever S., (2016), *Moral Particularism and Moral Generalism*, in "The Stanford Encyclopedia of Philosophy", N. Zalta E. N. (a cura di), 2016. Disponible à l'adresse: https://plato.stanford.edu/archives/win2020/entries/moral-particularism-generalism/
Rodogno R., Nørskov M., *The automation of ethics: the case of self driving cars,* in Hasse C., Søndergaard D. M., (a cura di) "Designing Robots – Designing Humans", Routledge, Paris (à paraître).
Ross W.D., *The right and the good*. Oxford: Clarendon Press, Oxford 1930.
Russell S.J., *Human Compatible: AI and the Problem of Control*. Allen Lane, Londres 2019.
Tappl R. (a cura di), *A Construction Manual for Robots' Ethical Systems*, Springer, New York 2015.
Wallach W., Allen C., *Moral machines: teaching robots right from wrong*, Oxford University Press, Oxford 2008.
Zagzebski L., *Exemplarist virtue theory*, in "Metaphilosophy", 41, 1-2, 2010, pp. 41–57.