MENSUEL N° 118
Avril 2018

philosophie magazine

MUHAMMAD YUNUS
Cet homme va-t-il rendre le capitalisme altruiste ?

REPORTAGE
KIBBOUTZ, LA DERNIÈRE UTOPIE ?

ISABELLE SORENTE
"MARC AURÈLE AIDE À SURMONTER LES CRISES EXISTENTIELLES"

CAHIER CENTRAL
Esprit de géométrie, esprit de finesse
PASCAL

Face à l'intelligence artificielle
COMMENT SAUVER L'INTELLIGENCE NATURELLE ?

Mensuel / France : 5,90 € Bel./Lux./Port. cont. : 6,50 € Suisse : 11 CHF Andorre : 6,20 € Allemagne : 6,90 € Canada : 11,50 $CA DOM : 8 € COM : 1 000 XPF Maroc : 70 DH Tunisie 11,30 TND

M 09521 - 118 - F: 5,90 € - RD

Les machines vont-elles nous faire la morale ?

Voitures autonomes, drones, hôpitaux: voilà des domaines dans lesquels les IA sont amenées à faire morts et merveilles. Mais peuvent-elles être programmées pour distinguer le bien et le mal ? Réponses de cinq spécialistes.

Par **Alexandre Lacroix**

Treize. Un chiffre qui porte malheur et qu'évitent les ascenseurs aux États-Unis. Pourtant, c'est le nombre de questions du test de la « Machine morale »[1], mis en ligne par le prestigieux Massachusetts Institute of Technology (MIT). Ce jeu, lancé en juin 2016, a recueilli près de 40 millions de réponses, un record. Le but est de sonder nos intuitions morales, quand une voiture autonome – du type Google Car –, fait face à une alternative dramatique. Admettons qu'il n'y ait que deux options : précipiter le véhicule contre un mur en tuant le conducteur, ou renverser trois enfants qui sortent de l'école, que faire ? Et si les enfants traversent la route alors que le feu est au rouge pour les piétons ? Et s'il y a aussi un bébé à bord de la voiture ?

Ce test n'est pas une simple affabulation : l'arrivée imminente sur les routes des

voitures autonomes nous oblige à résoudre ce qu'on appelle, dans la tradition éthique anglo-américaine, le « dilemme du tramway » [*voir les schémas ci-dessous*], formulé par Philippa Foot en 1967. Un tramway lancé à vive allure va percuter cinq personnes ; toutefois, il est en votre pouvoir de le dévier en actionnant un levier, auquel cas le tramway ne tuera qu'une personne. Le faites-vous, quitte à endosser la responsabilité de cette mort ? La plupart des gens répondent par l'affirmative. Maintenant, considérez cette variante : vous pouvez arrêter le tramway en poussant un homme obèse sur les rails. Le faites-vous ? La plupart des gens s'y refusent, même si le résultat est le même. Ce qui prouve que nous ne sommes pas à 100 % « conséquentialistes », c'est-à-dire que nous ne jugeons pas de la valeur morale d'une action qu'en fonction de son résultat sec.

BABY YOU CAN'T DRIVE MY CAR

L'un des principaux concepteurs du test du MIT, Jean-François Bonnefon, est chercheur en psychologie cognitive et professeur à la Toulouse School of Economics. Je le contacte et lui pose une question : si demain Google met en vente un véhicule programmé, dans certains cas, pour se crasher contre un mur, je serai très réticent à l'idée de monter à bord et n'y mettrai jamais mes enfants. Les gens ne passeront-ils à la voiture autonome que si elle sauve toujours son ou ses passagers, même quand c'est immoral ? « *Vous semblez préférer conduire seul. Mais admettons que la voiture autonome divise par dix, par cinq ou même par deux le risque d'accident, n'y a-t-il pas une certaine irrationalité à refuser ce bénéfice parce que vous craignez une situation rarissime ? Au contraire, n'avez-vous pas le devoir moral d'épargner à vos enfants les dangers inhérents à* votre *conduite ?* »

Serais-je donc le seul à redouter de me laisser transporter par un véhicule capable de me tuer pour sauver deux vies ? Je pose la question à Nicholas Evans, professeur de philosophie à l'université Lowell (Massachusetts) ; il vient, avec un groupe de chercheurs, de recevoir une bourse de 556 000 dollars de la National Science Foundation pour écrire un algorithme qui résoudra le dilemme du tramway. Je lui fais remarquer que le succès commercial des SUV, surtout auprès des pères de famille, est dû au fait que ces gros véhicules écraseraient à peu près n'importe quel obstacle. Ce n'est pas très altruiste, peut-être, mais très rassurant. « *Voilà un argument amusant !* commente Evans. *Supposez que vous soyez utilitariste et conséquentialiste. Vous voulez maximiser le bien-être du plus grand nombre, n'est-ce pas ? Vous pensez donc qu'un monde où toutes les personnes se déplaceraient en voiture autonome est souhaitable, puisqu'il y aurait à peu près 90 % d'accidents de voiture en moins. Seulement, vous savez qu'un tel monde n'existera que si ces voitures autonomes protègent dans tous les cas leurs passagers, sinon les gens refuseront de monter à bord. Un utilitariste cohérent acceptera qu'on adopte ici une règle déontologique non négociable – les voitures protégeront toujours leurs passagers – afin de maximiser le bien-être collectif. C'est ce qu'on appelle le* self-effacing utilitarianism, *l'"utilitarisme qui s'efface". Il est parfois rationnel, pour atteindre un but strictement utilitariste, d'accepter une règle déontologique !* »

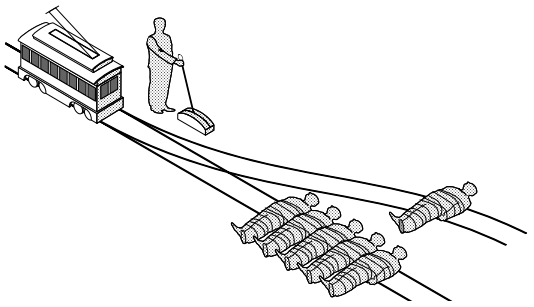
Mais Jean-François Bonnefon ne l'entend pas de cette oreille : « *En tant que psychologue, je pense que nous avons à combattre une hypocrisie fondamentale. Certaines décisions morales sont douloureuses à prendre, nous n'avons pas envie d'y penser. C'est pourquoi nous défendons l'idée qu'il y a de l'aléa, que cela fait partie de la condition humaine. Avec un peu de lâcheté, nous nous disons : "N'adoptons pas une approche systématique ! Si une situation extrême se présente, on verra bien, selon l'inspiration du moment." Le problème que posent les voitures autonomes – et que je vois, quant à moi, comme une opportunité – est qu'elles nous obligent à tout mettre à plat. Nous allons devoir nous entendre sur des principes, repérer les cas où ces principes ne sont pas satisfaisants, les discuter et trancher. Nous voici donc obligés d'entrer dans la zone d'ombre, que nous étions soulagés jusqu'ici d'entretenir. Et notez bien que cela ne détruit pas votre libre arbitre ! Supposez que vous ayez la possibilité, chez vous, assis dans un fauteuil, de décider ce que vous souhaitez que votre voiture fasse en situation d'urgence. Avec la voiture autonome, tout se passera comme si vous ne pouviez plus revenir sur un tel choix, une fois posé.* » Sauf que, objecté-je, une décision ne peut être prise qu'en contexte. Admettons que je me rende au chevet de ma femme mourante et que je sois l'unique soutien financier de mes enfants, n'ai-je pas une bonne raison de refuser de me sacrifier, quitte à faucher trois ou quatre personnes ? « *Vous devez envisager tous les cas de figure dès à présent, même celui-ci, et décider. Nous sommes en train de passer d'une situation de flexibilité ou plus exactement d'hypocrisie, où il nous est toujours loisible d'avancer les meilleures raisons du monde pour justifier nos arrangements avec la morale* a posteriori, *à une décision éthique* a priori. »

Élucider la zone d'ombre. Est-ce possible ? Nicholas Evans, qui se chamaille avec son chat de l'autre côté de l'Atlantique tandis que nous conversons sur Skype, est moins radical : « *Le problème est que nous ignorons ce que serait une voiture vraiment déontologique. Prenons l'un des commandements moraux les plus universels : "Tu ne tueras point." Peut-on imaginer une voiture qui respecterait ce commandement ? Non ! Le simple fait que nous programmions une machine nous contraint donc à nous placer dans une perspective conséquentialiste et utilitariste, même si ces options philosophiques nous répugnent.* »

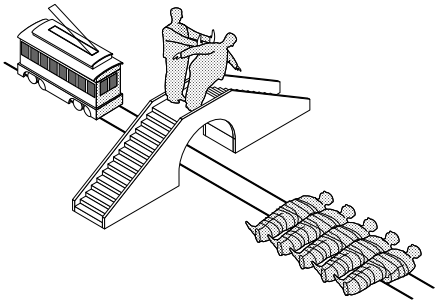
Cela me fait songer à un autre paradoxe : si les voitures autonomes s'imposent, il n'y aura plus de frein à la consommation d'alcool, puisque plus personne n'aura à prendre le

LE DILEMME DU TRAMWAY

01 **En 1967, la philosophe Philippa Foot a inventé ce dilemme :** un tramway lancé à pleine vitesse menace d'écraser cinq personnes sur la voie. Actionnez-vous le levier d'aiguillage qui permet de ne tuer qu'une seule personne ?



02 **Variante proposée par la philosophe Judith Jarvis Thomson en 1985 :** vous pouvez bloquer le tramway en poussant sur la voie un homme obèse, tuant celui-ci, mais sauvant les autres. Que faites-vous ?



volant. « *Oui*, s'amuse Evans, *mais il est rare que les alcooliques boivent au point d'en mourir, ce qui signifie que les vies humaines sauvées sur la route resteront excédentaires.* » Un autre problème me vient à l'esprit : les dons d'organes. Les accidentés en état de mort cérébrale sont aujourd'hui la principale source d'organes à transplanter. Que ferons-nous sans cet approvisionnement ? « *Une réponse possible est que nous devons faire avancer la recherche sur la fabrication d'organes de synthèse plus vite que les voitures autonomes.* »

Soit, mais revenons au dilemme du tramway : devrait-on sauver les jeunes davantage que les vieux, c'est-à-dire non pas sauver le maximum de vies, mais le maximum d'années de vie ? « *Un parallèle est ici possible avec l'allocation des organes pour les greffes,* avance Evans. *Même si c'est un sujet pénible, on peut soutenir que l'énergie que la société a investie dans un très jeune enfant n'est pas grande. C'est pourquoi les personnes prioritaires pour les transplantations ne sont pas les plus jeunes, mais les adolescents. La tranche 15-25 ans sera donc considérée comme prioritaire, puis viendront les enfants, puis les plus vieux. Actuellement, les voitures ne connaissent pas votre âge. Mais si vous pilotez une Google Car et que vous avez un téléphone fonctionnant sur Android, tout comme les piétons ou les autres passagers en face de vous, on peut imaginer que ces données soient disponibles et qu'on applique ce critère de répartition.* » Néanmoins, j'entrevois ici un énorme problème : je suppose qu'il existera toujours, dans l'avenir, des voitures de luxe, comme des Porsche ou des Maserati. Or ce n'est pas entre 15 et 25 ans qu'un conducteur peut consacrer le plus gros de son budget à l'achat d'un véhicule, mais plutôt après 50 ans. Les voitures de luxe seront-elles jeunistes, ou donneront-elles la priorité à leurs propriétaires d'âge mûr ? « *C'est malheureusement vrai*, reconnaît Nicholas Evans, *les constructeurs de véhicules très chers pourraient être amenés à favoriser leur cœur de marché, c'est-à-dire des gens en moyenne assez âgés ou bénéficiant d'un statut social privilégié. Je ne vois qu'un frein régulateur à cette tentation : la mauvaise publicité. "La limousine de la marque X est une tueuse de bébés et d'enfants" : je vois déjà les gros titres dans la presse !* »

Autre question : qui est responsable en cas d'accident ? Celui qui a écrit l'algorithme, vous-même Nicholas Evans, soit dit sans vouloir vous offenser ? « *Non, je ne fais pas un algorithme pour des voitures réelles, mais pour des simulations ! Cependant, la voiture autonome amène à repenser pas mal de choses. La notion de conducteur va disparaître, n'est-ce pas ? Il n'y aura donc plus d'assurance conducteur ! Immense économie pour les foyers. Par contre, il y aura encore des accidents, même s'ils seront plus rares. Aujourd'hui, la responsabilité d'une mort n'est pas imputée à Seat, Renault ou Volkswagen. Les voitures tuent, mais les constructeurs sont hors de cause. Il faudra sans doute imaginer une contribution par l'impôt, moins importante que l'assurance, qui permettra à l'État de "réparer" les victimes de tous les accidents de la route. Car les acteurs privés – les constructeurs – pourraient refuser de construire ces véhicules par crainte des procès. Donc, il est bien possible que, si la voiture est aux normes, qu'elle a été testée et que son algorithme est conforme aux réglementations en vigueur, ce soit l'État qui couvre le risque.* »

> « Peut-on imaginer une voiture qui respecterait le commandement 'Tu ne tueras point' ? »
>
> **NICHOLAS EVANS,** PHILOSOPHE

Firefly de la marque Waymo est un prototype fonctionnel de voiture autonome développé pour Google.

EN ATTENDANT LES ROBOTS TUEURS

À la différence des voitures autonomes, il existe des machines conçues exprès pour tuer : les drones. Depuis quelques années, l'industrie de l'armement – notamment aux États-Unis – se propose de mettre au point des « robots tueurs », soit des drones autonomes qui ne seraient plus pilotés à distance par un militaire, comme c'est le cas aujourd'hui pour les *Predator*, et qui suivraient, une fois déployés, leur propre ordre de mission.

L'un des principaux arguments en leur faveur, avancé par le roboticien Ronald C. Arkin, est qu'ils seraient susceptibles de s'autodétruire, alors qu'un militaire humain, sur le terrain, voudra sauver sa peau coûte que coûte, quitte à tirer largement autour de lui. Par exemple, un drone imploserait plutôt que de risquer de tuer un civil ou un enfant. « *Je ne suis pas convaincu par cet argument*, explique Peter Asaro, professeur associé à la New School de New York et spécialiste en éthique

de la robotique. *D'un côté, il est difficile de prédire le comportement des humains sur un champ de bataille, et nous savons que beaucoup de soldats se sont sacrifiés par le passé pour sauver leurs camarades ou défendre une position. De l'autre, je ne vois guère d'exemples de commandements militaires qui auraient sacrifié du matériel technologique de pointe. Quelle armée a détruit des tanks, des hélicoptères ou des drones afin d'épargner des civils?* »

Peter Asaro est cofondateur de l'Icrac (*International Committee for Robot Arms Control*, « comité international pour le contrôle des armes robotiques »), une organisation non gouvernementale qui regroupe des chercheurs proposant des textes de lois afin d'encadrer l'usage des machines dans la guerre ; l'un des faits d'armes de l'organisation est la campagne « Interdisez les robots tueurs ! » de 2015, avec un appel à l'interdiction des drones autonomes signé par 1 500 personnalités, dont Elon Musk, Stephen Hawking ou le philosophe Daniel Dennett.

Or un autre argument des partisans des systèmes autonomes est qu'une machine peut être programmée pour ne jamais commettre de bavure, ne pas ouvrir le feu sur les civils, sur les femmes, les vieillards ou les enfants. « *L'un des principes fondamentaux du droit de la guerre est la distinction*, admet Peter Asaro, *c'est-à-dire la capacité à distinguer la population civile des combattants. Si cela reposait entièrement sur la reconnaissance visuelle, les machines seraient sans doute performantes dans ce domaine. Mais ce n'est pas aussi simple, à cause de la catégorie des "civils participant aux hostilités". Ici, le critère n'est pas visuel mais comportemental. Est-ce que tel civil ne fait que ramasser une pierre et la lancer sur les chars par colère ou est-il manifeste, à sa manière de se mouvoir sur le terrain, qu'il est en train d'exécuter un ordre d'un commandement militaire ? Représente-t-il une menace ? Ces aspects comportementaux ne peuvent être décryptés qu'avec un certain sens des interactions sociales, de la psychologie. Le droit de la guerre n'autorise à cibler que les civils participant* directement *au combat, or les machines ne savent pas les identifier.* »

Le second grand principe du droit de la guerre est la proportionnalité : on ne détruit pas une ville entière pour éliminer cinq personnes, autrement dit l'emploi de la force doit être proportionnel aux objectifs militaires poursuivis et aux menaces réelles. « *Ce principe de proportionnalité nécessite lui aussi un jugement humain. Les dilemmes sont parfois délicats : quelles infrastructures risquez-vous d'endommager pour atteindre votre objectif militaire ? Imaginons qu'un hôpital ou une école soient situés à côté d'un entrepôt de munitions, vous prenez le risque de bombarder ce dernier ? Si le risque est de frapper une école, vous pouvez bombarder la nuit ; mais cette solution ne vaut pas pour un hôpital. Il n'y a pas de réponse standard. Votre propre stratégie militaire, vos objectifs évoluent parfois d'heure en heure. C'est pourquoi un commandement militaire humain doit endosser ces décisions et prendre la responsabilité légale de l'ordre donné. Si c'est une machine qui a tranché et qu'une école avec trois cents enfants a été bombardée, qui est responsable de ce massacre ? La machine ? Le programmeur ? Nous n'avons pas le choix : il faut, comme le veut la doctrine pour l'instant dominante de l'armée américaine,* "garder des humains dans la boucle". »

C'est aussi le point de vue de Nolen Gertz, professeur de philosophie à l'université de Twente (Pays-Bas), autre auteur de référence sur ce thème : « *La tradition de la "guerre juste", qui remonte à saint Augustin, nous enseigne que, moralement, il est mauvais que des combattants prennent plaisir à tuer. Il n'est de guerre juste possible que si les acteurs ne sont ni des nihilistes ni des sadiques, que foncièrement ils détestent tuer, mais qu'ils vont à l'encontre de leurs principes moraux parce qu'ils s'estiment dans un cas de légitime défense ou de force majeure. C'est une posture subtile : la guerre n'est juste que si ceux qui la mènent la font malgré eux. Cette tension disparaît avec des machines.* »

Certes, mais l'argument est réversible : si jamais vous disposez de robots capables de mener complètement une guerre à la place des humains, et que vous avez donc la possibilité de l'emporter sans exposer la vie d'aucun de vos militaires et sans même que les décisions létales viennent peser sur les consciences des pilotes de drones, pourquoi n'éviteriez-vous pas toutes les souffrances à votre population ? « *Ce raisonnement ne me paraît pas recevable,* maintient Gertz, *pour cette raison que le rôle de la technologie n'est pas simplement instrumental, neutre. La technologie affecte, forme et influence nos comportements,*

> « La guerre n'est juste que si ceux qui la mènent la font malgré eux. Cette tension disparaît avec des machines »
>
> **NOLEN GERTZ,** PHILOSOPHE

Un soldat français participant à l'opération Barkhane au Niger veille sur un drone *MQ-9 Reaper* conçu pour la surveillance et le combat.

© Pascal Guyot/AFP

> Lors d'une opération délicate, la sonde tombe en panne, avec des dommages pour le patient. Qui est responsable?
>
> **KENNETH W. GOODMAN,** PROFESSEUR DE MÉDECINE

Une équipe médicale de l'hôpital de Chattanooga (États-Unis) assistée par le robot Da Vinci s'apprête à pratiquer une ablation de la prostate.

© Dan Henry/AP/Sipa

ainsi que les processus de décision éthique. Imaginons que, comme vous le dites, le fait de déclarer la guerre puis de décimer un adversaire n'entraîne aucun coût humain, ni même psychologique, que ce soient des robots qui fassent le sale boulot. Cela abaisse vertigineusement le seuil d'acceptabilité de la guerre pour une opinion publique, un pouvoir exécutif ou un état-major. En d'autres termes, cela rend beaucoup trop facile l'acte de tuer. Imaginez un monde où prolifèrent des robots ayant un permis de tuer. Ce monde vous paraît-il souhaitable? »

ÉLÉMENTAIRE, MON CHER WATSON?

Mais il est un troisième domaine à haute implication éthique où arrive rapidement l'IA: la santé. Watson, l'ordinateur d'IBM qui a remporté un championnat de *Jeopardy*, a récemment diagnostiqué une forme rare de leucémie à une patiente, à l'université de Tokyo, en consultant vingt millions d'articles de recherche en dix minutes; l'ordinateur a par ailleurs proposé un traitement adapté. Mais supposons qu'une IA commette une erreur grave de diagnostic: qui sera responsable? « *Voilà un magnifique problème d'agentivité*, démarre avec enthousiasme Kenneth W. Goodman, professeur de médecine et directeur de l'Institut de bioéthique de l'université de Miami. *Prenez le cas d'un chirurgien qui utilise, depuis dix ans, une sonde avec succès; lors d'une opération délicate, la sonde tombe en panne, avec des dommages pour le patient. Le chirurgien est-il responsable? Oui, si vous considérez qu'il a le même statut que le capitaine d'un bateau et qu'il doit assumer le cours des événements. Mais, à mon sens, il est plutôt question d'une responsabilité partagée. Ainsi les systèmes d'aide à la décision clinique* [SADC] *qui se multiplient nous font entrer dans une ère où il y a beaucoup de niveaux de responsabilité. Qui a construit la banque de données consultée par l'IA? Qui a écrit le programme? Question d'autant plus difficile que la plupart des codes utilisés ont été écrits par plusieurs personnes, certains proviennent en partie de logiciels* open source, *transformés par des chercheurs ou des sociétés pour une application précise. Nous allons devoir poser des normes internationales dans ce no man's land.* »

Quant à Peter Asaro, qui s'est aussi intéressé à la question de la santé, il a cette remarque complémentaire, bien qu'un peu glaçante: « *Les hôpitaux sont confrontés à des dilemmes éthiques liés au triage et à l'allocation de ressources rares, comme le sang à transfuser ou les organes à transplanter. Quels patients vont en bénéficier? Voilà un cas où la subjectivité des médecins, leur sympathie pour certains patients ou même la somme d'argent qu'un patient est disposé à payer pour ces soins vitaux, risquent de biaiser l'allocation. Dans ce domaine très précis, il est possible que les recommandations d'une IA soient plus objectives, meilleures éthiquement, même si elles devraient être soumises à l'approbation d'un docteur.* »

C'est l'un des cas, assez rares en fait apparaît-il à l'issue de cette enquête, où l'humain risque d'avoir un jugement moral moins pertinent que celui de l'ordinateur. Et pour le reste? « *Il y a une histoire que j'adore et qui remonte aux balbutiements de l'IA*, poursuit Kenneth Goodman. *Un robot a devant lui trois objets, une sphère, une pyramide et un cube en bois. Un examinateur lui demande de poser la pyramide sur le cube. Il y parvient. Bravo! L'examinateur lui ordonne ensuite de poser la sphère sur la pyramide. Il essaie, la sphère roule. Il essaie encore, et encore, et encore... Le robot ignore que c'est impossible. Cette histoire est une parabole utile pour notre temps: nos jugements moraux supposent un arrière-plan de connaissances implicites, qui ne sont pas toutes explicitables ni programmables, car elles sont innombrables. C'est ce que Wittgenstein appelle le domaine de la certitude.* » Oui, ça me revient, Wittgenstein donne d'ailleurs un exemple saisissant: un homme qui couperait le bras de son voisin pour voir s'il repousse ne serait pas un scientifique tentant une expérience, mais un fou. « *Exactement! C'est ça le problème! Nous autres humains ne nous contentons jamais de suivre une règle. Car les règles explicites ne valent que dans un rayon d'action très étroit. La différence entre l'IA et l'humain, ce n'est pas seulement l'interprétation du contexte, c'est que l'humain tient un nombre énorme de principes pour tellement acquis et évidents, qu'ils ne méritent même pas d'être énoncés. C'est pourquoi le jugement éthique humain reste indispensable.* » Faudrait-il ajouter: dans l'état actuel de nos technologies?