philopsis

La morale

L'éthique artificielle ou l'éthique d'après l'intelligence artificielle[1]

Laurent Cournarie

Philopsis : Revue numérique
http://www.philopsis.fr

L'intelligence artificielle (*IA*)[2] est en train de bouleverser nos vies sous nos yeux, malgré nous et avec notre consentement. Ce n'est pas la première révolution. Lévi-Strauss considérait qu'il y avait eu deux révolutions majeures dans l'histoire de l'humanité : la révolution néolithique et la révolution industrielle. Il se pourrait que la révolution numérique soit la troisième du genre. Elle en possède la radicalité, mais avec pour spécificité d'être une révolution de et par l'intelligence. Révolution par l'intelligence — on est passé à une société de la connaissance, c'est-à-dire à une société où tous les rapports sont médiatisés par des systèmes informatiques ; révolution de l'intelligence parce que l'intelligence est (dite) artificielle.

[1] Version augmentée pour *Philopsis* d'une conférence organisée et publiée par *NXU* : http://nxu-thinktank.com/colloque-nxu-5-juin-2018-lethique-d-apres-lia/.

[2] L'«intelligence artificielle » (*IA*) désigne la partie de l'informatique qui a pour objet la création d'un programme informatique capable de conférer à une machine dans laquelle il est implanté un comportement "intelligent". L'« éthique artificielle », ici prise pour synonyme des expressions « *robots ethics* » ou « *machine ethics* » plus couramment utilisées, est la partie de l'intelligence artificielle qui a pour objet la création d'un programme informatique capable de conférer à une machine dans laquelle il est implanté cette forme particulière de comportement intelligent qu'est un comportement moral, c'est-à-dire intégrant la polarité du bien et du mal.

Nous proposons cette expression : "l'éthique d'après l'*IA*" à la place de "l'éthique appliquée à l'*IA*" qui sert d'habitude à définir l'éthique artificielle pour suggérer que l'*IA* n'est pas seulement l'objet de l'éthique, mais aussi le sujet de l'éthique, que l'*IA* est susceptible de transformer l'éthique qui, par elle-même, n'est évidemment rien de stabilisé.

On le sait, l'*IA* suscite le mythe : soit l'utopie d'un monde où le travail pénible et répétitif sera délégué à des machines, où la maladie pourra être éradiquée, où le crime prédit, etc. — bref où le mal sera annulé ; soit l'apocalypse d'un monde ayant atteint le point de singularité où l'IA forte remplacera l'intelligence humaine (naturelle) et deviendra(it) potentiellement hostile à l'humanité — bref où le cauchemar sera devenu réalité.

L'avenir ne sera ni l'utopie d'une humanité libérée par l'*IA* de toutes ses servitudes, ni la dystopie d'une humanité asservie à l'*IA*, l'histoire étant, même et surtout technologiquement, imprévisible. Mais il n'en demeure pas moins que l'*IA* pose, au-delà des défis techniques, de multiples questions à la fois sociétales, juridiques et éthiques.

Le thème « *IA* et éthique » est très débattu et déjà rebattu. Certains peuvent considérer que l'éthique en *IA* est un thème marketing (*AI for Humanity*) et, qu'à s'en faire le spécialiste, on risque de manquer le développement scientifique et économique de et par l'*IA* : le gagnant en éthique serait le perdant en économie. L'éthique de la responsabilité, pour ainsi dire, recommanderait de surmonter la réflexion éthique sur l'*IA* pour se lancer dans le financement et la recherche en *IA* : la recherche de l'éthique en *IA* ne serait que le supplément d'âme de la recherche scientifique et industrielle. D'autres au contraire, considérant, au nom d'une sorte d'éthique de la conviction, que certaines valeurs éthiques sont essentielles à l'humanité, que l'*IA* doit augmenter l'homme, non le remplacer, s'élèvent contre toute naïveté à croire que les algorithmes sont neutres, qu'ils ne recèlent aucun biais cognitif, que leurs résultats en *deep learning* sont transparents et explicables. L'éthique ici se présente comme une intelligence critique (des dérives) de l'*IA*. Donc on s'inquiète[3] ainsi des effets éthiques de l'*IA*, en insistant sur la nécessité : (a) de protéger les données personnelles — l'éthique commence par le respect de la liberté individuelle ; (b) de lutter contre les discriminations — l'éthique repose aussi sur le principe d'égalité des personnes[4] ; (c) de préserver le pluralisme humain contre une tendance à la « clusterisation » des individus — l'éthique suppose le pouvoir pour l'individu d'être le sujet propre de sa vie propre ; (d) la nécessité de défendre le bien commun ou l'intérêt général des sociétés et peut-être de l'humanité en général contre d'une part des entreprises privées et d'autre part contre l'individualisme de l'usager et du consommateur — l'éthique est aussi une question politique.

Mais plus précisément comment l'*IA* interroge-t-elle l'éthique ou comment l'éthique peut-elle se saisir de l'*IA*, si l'on ne confond pas tout ce qui n'est pas technique avec l'éthique et l'éthique avec le sociétal ? Que fait l'*IA* à l'éthique ? L'éthique est à la mode et l'*IA* s'impose à la réflexion collective. Alors comment traiter du rapport entre éthique et *IA* ? On peut avancer trois idées simples : (1) le monde technoscientifique contemporain n'est pas moins éthique mais plus ; (2) cette croissance éthique est peut-être en même temps une crise morale ; (3) l'enjeu éthique de l'*IA* se concentre sur l'autonomie humaine de la décision.

[3] Cf. le rapport de la CNIL (décembre 2017) sur Les enjeux éthiques des algorithmes et de l'intelligence artificielle.

[4] Entre autres exemples, le programme *IA* d'*Echo look*, un coach stylistique lancé par Amazone, avait éliminé les candidats de couleur noir, ayant déduit à partir des données fournies, que la peau claire était un critère de beauté.

(1) Le monde est plus éthique sous deux aspects. Il l'est d'abord « sociétalement ». L'éthique est à la mode et elle est partout : éthique des affaires, éthique des entreprises, bioéthique, éthique environnementale, éthique animale, etc. Tout est éthique ou tout socialement pose un problème éthique. Mais il l'est socialement parce qu'il l'est surtout problématiquement. Relève désormais de l'éthique ce qui jusque-là n'en avait jamais relevé. L'éthique désignait et désignait toute théorie normative de l'action : comment bien agir ou que doit-on vouloir ? — étant admis que toutes les actions ne se valent pas, que les sociétés valorisent certaines comme bonnes ou justes et d'autres comme mauvaises ou injustes. Il n'y a pas d'éthique ou de morale en deçà de cette distinction entre du bon et du mauvais et du désir ou de la volonté d'orienter l'action en fonction de ces valeurs. Ce qui a changé c'est, pour ainsi dire, le périmètre de l'éthique ou de la morale. Jusqu'à présent, la norme éthique concernait : 1. le rapport de l'homme à l'autre homme — étaient exclus de l'éthique la nature, l'animal, la machine ; 2. dans les limites du temps présent, au moins possible (devoirs envers le prochain) — étaient exclus de l'éthique les générations à venir ; 3. et plus fondamentalement, dans ce qu'on peut appeler les limites de la finitude humaine — étaient exclus de l'éthique tout ce qui touche aux fins de l'existence (naissance/mort)[5]. Or ce que nous subissions, nous sommes en passe de pouvoir le vouloir : l'impossible devient possible. *From chance to choice*. La nature et le hasard ont longtemps décidé de la naissance, de la vie et de la mort. L'extension du périmètre de l'éthique est ainsi en réalité l'augmentation (du cercle) de la responsabilité humaine. Le monde est donc plus éthique parce que nous devons collectivement normer nos actions sur la naissance, la vie et la mort qui avaient, au moins dans leur possibilité, toujours échappé au champ de la responsabilité morale.

(2) Mais en même temps le monde est peut-être moins moral. On assiste à un recul de la capacité à poser une loi catégorique ou à interdire inconditionnellement une action pragmatiquement ou techniquement possible (loi de Gabor). L'humanité vit en quelque sorte dans un autre monde éthique. L'ancien monde était fait de devoirs et d'interdits, investis et cautionnés par des institutions. Or, pour des raisons multiples[6], il est de moins en moins possible d'imposer une règle pour interdire un projet, voire seulement de prolonger un moratoire. Le « non » s'efface de l'horizon notre monde. Les intérêts économiques, la disparité des systèmes juridiques, le désir d'apprendre et de connaître davantage, la revendication des droits et de l'égalité et, tout simplement, ce qu'il est convenu d'appeler l'évolution des mœurs, sont plus puissants que les scrupules moraux ancestraux.

[5] Descartes, au XVIIème siècle, écrivait que « s'il est possible de trouver quelque moyen qui rende communément les hommes plus sages et plus habiles qu'ils n'ont été jusqu'ici, je crois que c'est dans la médecine qu'on doit le chercher » (*Discours de la méthode*, VI). Mais désespérant des progrès de la médecine à laquelle il avait consacré beaucoup d'efforts, il estime finalement dans une *Lettre à Chanut* (15 juin 1646) qu'« au lieu de trouver les moyens de conserver la vie, j'en ai trouvé un autre, bien plus aisé et plus sûr, qui est de ne pas craindre la mort ». Autrement dit la maîtrise de soi, de ses désirs et de ses craintes, l'effort pour acquérir des vertus (courage, tempérance, justice…) au fondement de l'exigence éthique sont peut-être la seule réponse que l'humanité pouvait se donner tant qu'elle était impuissante à peser sur le destin de sa vie biologique.

[6] Recul de l'autorité des institutions, individualisme, pluralisme des valeurs dans des sociétés de plus en plus multiculturalistes, compétition mondiale, etc.

C'est à l'aune de ce double contexte (inflation éthique, déflation morale) qu'il faut situer l'enjeu éthique de l'*IA*. Pourtant l'*IA* pose des problèmes éthiques spécifiques. Que peut-être l'éthique « d'après » l'*IA* ? En quel sens peut-on parler d'une éthique artificielle ? Car il ne s'agit plus simplement de savoir comment l'homme doit user des machines, il s'agit de savoir comment il doit programmer les machines pour un comportement éthique intelligent. L'éthique passe de l'usage humain de la machine (et plus généralement de la technologie) à la capacité d'action autonome de la machine. Autrement dit, l'éthique artificielle ne désigne pas seulement l'éthique appliquée à l'*IA* (l'*IA* comme objet de l'éthique[7]) mais l'*IA* comme sujet de l'éthique.

(3) L'enjeu éthique de l'*IA* se concentre sur l'autonomie humaine de la décision. L'*IA* ne menace-t-elle pas le pouvoir de l'individu (mais aussi des groupes) de décider lui-même de sa vie, de maîtriser sa décision ?

Une bonne illustration de cet enjeu est le cas de l'automobile autonome. Il ne s'agit pas seulement un véhicule sans conducteur embarqué — un drone est de ce type, mais il est piloté à distance, sorte de « télé-véhicule »[8]. L'automobile autonome est un robot : c'est encore (extérieurement) une voiture, mais dotée d'une *IA*, constituée de milliers de lignes de programme informatique articulées à des capteurs multiples qui lui permettent d'agir et de réagir à un environnement en mouvement. Si agir consiste à produire un mouvement intentionnel dont l'agent contrôle en permanence l'adéquation avec le but recherché, alors une "autonomobile", comme on se propose de l'appeler, est définissable comme un « agent pratique artificiel modulaire » (*APAM*) — mais c'est également vrai d'un robot aide-soignant : (a) un agent « pratique » parce qu'elle poursuit de manière intelligente un but pratique en s'adaptant à ses objectifs (feux de signalisation, limitation de vitesse, passage de piétons, circulation…) et prenant d'elle-même des décisions pratiques appropriées (ralentir, freiner, s'arrêter, changer d'itinéraire…) ; (b) un agent pratique « artificiel » parce qu'elle est une *IA* (programme informatique qui commande des séries d'impulsions électroniques) ; (c) un agent pratique artificiel « modulaire » parce qu'elle n'est capable, contrairement à l'agent humain, que d'un spectre restreint de buts pratiques[9].

Un *APAM* n'est pas un robot de science-fiction (l'objet fictionnel d'un monde fictionnel) mais l'objet intelligent qui, dans un avenir proche, évoluera dans notre monde social quotidien, impliquant des humains et donc des dommages pour eux. Un *APAM* d'une part dispose d'une marge d'initiative (il prend des décisions par rapport à des buts pratiques définis) et, d'autre part est susceptible, du fait de son action et de son autonomie de décision, de

[7] L'*IA* interroge l'éthique de manière très variée, on l'a vu : opacité des processus de prise de décision (le problème de la boîte noire, Knight, 2017), la fracture entre les pays dans le développement de l'*IA*, la juste répartition de la croissance de la productivité par l'*IA*, la surveillance et la discrimination invisibles : cf. Martin Gibert, « L'éthique artificielle », *l'Encyclopédie philosophique* : http://encyclo-philo.fr/etique-artificielle-gp/.

[8] Pour ce qui suit, notamment sur le concept d'*APAM*, nous nous référons aux réflexions de Stéphane Chauvier dans son article « L'éthique artificielle », *l'Encyclopédie philosophique* : http://encyclo-philo.fr/ethique-artificielle-a/

[9] Par exemple, *lethal autonomous weapons systems*, ou robots aides-soignants, voire robots d'assistance sexuelle cf. D. Levy, *Love and Sex with Robots*, New York, Harper Collins, 2007.

nuire aux humains. C'est pourquoi, la mise en circulation d'*APAM* ne peut se contenter de la limitation habituelle du risque technologique. Il s'agit d'agir sur les procédures décisionnelles de l'*APAM*, ce qui implique d'intégrer dans l'IA une éthique artificielle, c'est-à-dire des lignes de programme supplémentaires garantissant un comportement conforme ou compatible avec les normes juridiques, les intérêts humains, les valeurs sociales. Mais un *APAM* peut-il être et devenir un agent moral artificiel modulaire (*AMAM*) ? Et d'abord, quelle éthique implémenter dans un *APAM* et en particulier dans une autonomobile ?

Le problème éthique de l'automobile autonome concerne la programmation d'une machine à choisir en situation critique. Ce n'est plus l'homme (l'individu) qui choisit, mais l'homme (l'ingénieur) qui doit choisir comment la machine doit choisir à la place de l'individu : comment programmer la machine à choisir qui dispensera l'homme à l'avenir de choisir.

Mais ce problème éthique reste singulier, y compris par rapport à l'application de l'*IA* à d'autres domaines (juridique, chirurgical), car il s'agit de confier à un programme-machine une décision que le conducteur humain ne prend jamais de manière réfléchie. En situation critique, un conducteur réagit comme il peut, par réflexe en cherchant à éviter un obstacle et à sauver sa vie, sans que la priorité entre les deux objectifs soit soumise à l'examen. L'action ne souffre pas de délai et donc ne laisse pas le loisir au conducteur de choisir entre plusieurs options éthiques (éviter de nuire à autrui, même au prix de sa vie, ou sauver sa vie, même si c'est au détriment de l'intégrité d'autrui). L'accident se présente toujours à l'agent moral (pourvu de certains principes, éduqué à certaines valeurs), comme un événement, que chacun négocie comme il peut en situation, et non sous la forme d'un dilemme. Or dès lors qu'on passe à une voiture autonome, il faut déterminer les règles que le véhicule doit suivre en situation critique. Entre éviter une personne (ou un groupe de personnes) ou préserver sa vie : l'*IA* doit être programmée à choisir l'un ou l'autre — tandis que l'homme peut choisir l'un en se reprochant de n'avoir pas choisi l'autre, ou refuser de choisir — choisir de ne pas choisir au motif que le choix est impossible, insignifiant ou que le contexte du choix est lui-même immoral. Mais ici la possibilité de choisir en ayant conscience de mal choisir ou de choisir de ne pas choisir est indisponible. La conscience morale ou la mauvaise conscience fait défaut au choix.

On peut alors concevoir deux types d'autonomobiles : un véhicule exclusivement altruiste (choisissant de sauver systématiquement les piétons par exemple, même au risque de la mort du conducteur et des passagers) ou égoïste (choisissant de sauver systématiquement le conducteur et les passagers au risque de la mort des piétons). Les enquêtes d'opinion montrent la préférence forte pour un véhicule qui assure en priorité la sécurité de leur utilisateur et de leur propriétaire — sinon la voiture deviendrait une machine sacrificielle, peut-être alors superlativement éthique, mais contre son usager, ce qui en fait un argument de vente douteux. Pour programmer l'*IA* éthiquement, on accumule les données à partir des réponses à plusieurs scénarios pré-définis[10]. On propose, dans le cas où une collision est inévitable, de faire un choix : selon l'âge (enfants ou personnes âgés), la nature (êtres humains ou animaux), la position (passagers ou piétons), le nombre — plusieurs combinaisons étant possibles.

[10] Cf. la plateforme *Moral machine.*

Evidemment, le problème éthique ne concerne pas l'*IA* mais la programmation humaine de l'*IA*. A moins d'une *IA* forte, une décision n'a de sens moral que pour l'homme. L'*IA* n'agit pas éthiquement mais suit un algorithme programmé selon des normes éthiques pré-définies par l'homme. Mais que se passe-t-il si, malgré tout, l'*IA* est défaillante ? Il n'y a pas faute morale puisque l'*IA* n'agit pas moralement. On n'a affaire qu'à une erreur de programmation ou à une défaillance. Autrement dit, la valeur morale de l'action est remplacée par une valeur technique et, de toute façon, ne concerne plus l'individu en tant que tel, déchargé de sa responsabilité (au-delà du choix d'acheter ou non un véhicule autonome, et de son comportement dans le véhicule (travailler, consommer de l'alcool[11]). La faute ne peut concerner éventuellement que le délinquant qui pirate l'*IA* (ce qui est toujours possible) ou l'ingénieur ou le constructeur — mais alors on verse dans le droit.

Mais cette déresponsabilisation de l'individu qu'on pourrait dénoncer n'est peut-être qu'apparente. En effet, ce qui sera éthique se déplace en amont vers le choix d'acheter une voiture autonome plutôt qu'une voiture conventionnelle, s'il est avéré qu'elle est plus sûre, comme les études tendent à le prouver. Ce sera un choix éthique, du point de vue utilitariste, puisque ce sera un choix qui maximisera le bien pour soi et pour autrui en augmentant la sécurité globale[12], ou parce que l'autonomobile pourra être le lieu et l'occasion d'un accomplissement de soi, si l'on adopte le point de vue d'une éthique des vertus. Néanmoins, cela revient à dire que l'*IA* affranchit l'individu d'une responsabilité éthique sans l'assumer elle-même pour autant : autrement dit, un dispositif technique se substitue à la responsabilité éthique individuelle. C'est si vrai que si on fait l'expérience de pensée d'un parc automobile intégralement composé de véhicules autonomes, on pourrait supposer qu'il n'y aurait plus (ou quasiment plus) d'accidents. Preuve donc que l'exigence éthique n'était que l'envers de la faillibilité humaine (tu dois parce que tu es faillible) : si on résout le problème de la faillibilité humaine, on résout le problème éthique en le faisant disparaître. Finalement, plus de dilemme moral si les *IA* interagissent entre elles en bonne intelligence artificielle. En bref, la promesse à l'horizon serait que la technique dispense à terme de l'éthique.

Alors que peut faire l'*IA* à l'éthique et que peut-être une éthique artificielle ? On peut conclure sur trois remarques.

1) Une remarque générale : l'éthique change avec la puissance technique. On peut décrire ce changement comme le passage de la morale à l'éthique, ou de l'éthique impérative à une éthique adaptative, mais l'idée (classique) que la technique serait par définition éthiquement neutre est problématique. Le dispositif technique induit des modes de penser et, en l'occurrence, d'évaluer. Le rapport éthique-technique change avec la technoscience. La technoscience n'est pas seulement un nouvel âge de la tech-

[11] Est-ce que la société sera plus morale ou immorale si l'autonomobile devient un bar ambulant ou un club de rencontres pour adultes consentants ?

[12] Selon les projections les plus optimistes, le déploiement de l'automobile autonome à large échelle permettrait d'éviter 90 % des accidents qui sont dû à une erreur humaine : conducteur fatigué, sous l'emprise de l'alcool ou d'une drogue, qui ne respecte pas le code de la route. Cf. «Algorithm Aversion : People Erroneously Avoid Algorithms After Seeing Them Err », *Journal of Experimental Psychology,* 2014.

nique, mais une nouvelle époque de la relation entre l'éthique et la technique — ce qui mérite réflexion.

2) Quelle éthique est programmable dans une *IA* ? Comment une éthique artificielle est-elle possible [13] ? Il est facile de comprendre que le problème éthique en *IA* prend la forme nécessaire du dilemme. Or le dilemme est posé en termes d'acceptabilité qui est le critère de l'éthique utilitariste. Est moral ce qui est acceptable parce qu'il constitue la plus grande somme de biens ou la moins grande somme de maux. Or l'utilitarisme (qui est la version la plus connue de ce qu'on nomme le conséquentialisme) est une des trois principales théories normatives. Donc l'*IA* ne rencontre l'éthique qu'en sélectionnant une théorie normative[14] — pour une raison évidente : c'est la seule qui prétende à une quantification (+ de maux/- de maux). Impossible d'introduire dans un algorithme la règle d'Or ou « tu ne tueras point » — l'éthique de l'*IA* n'est concernée que par « lequel et combien entre A ou B ? » Non seulement l'*IA* réduit l'éthique au dilemme en situation critique, mais le dilemme n'a pas de sens éthique pour l'*IA*.

3) on peut être partagé sur l'interprétation à donner de l'éthique d'après l'*IA*. On peut y voir l'occasion de clarifier nos intuitions éthiques et donc d'un progrès moral[15]. Car, somme toute, l'interdit du meurtre (inprogrammable) n'a jamais empêché, malgré son inconditionnalité, aucun meurtre. Donc, à quoi bon la conscience morale (irréductible à l'éthique artificielle) si elle n'empêche pas ce qu'elle interdit ? La programmation éthique de l'*IA* pointe une sorte d'hypocrisie ou de faiblesse pratique : nous n'agissons pas comme nous nous imposons normativement tous d'agir. En situation critique, il est probable d'ailleurs que les individus font le choix de ce qui maximise le bien ou minimise le mal, et que ce choix leur paraît raisonnable et moral.

Mais, d'un autre côté, la réduction de l'évaluation morale à une forme quelconque de calcul utilitariste peut apparaître comme un appauvrissement de l'expérience éthique dans sa richesse et sa spécificité humaine. En effet, l'expérience morale ne se résume pas au dilemme[16]. Si l'on délègue à l'*IA*

[13] Ethique artificielle (*EA*) ou *Machine ethics* : partie de l'intelligence artificielle qui a pour objet de créer un programme informatique capable de conférer à une machine dans laquelle il est implanté cette forme particulière de comportement intelligent qu'est un comportement moral, c'est-à-dire sensible à la polarité du bien et du mal.

[14] En philosophie morale contemporaine, on distingue la méta-éthique, l'éthique normative et l'éthique appliquée.

[15] C'est ce que soutiennent explicitement Michael et Susan Anderson : « One needs to turn to the branch of philosophy that is concerned with ethics for insight into what is considered to be ethically acceptable behavior. It is a considerable challenge because, even among experts, ethics has not been completely codified. It is a field that is still evolving. We shall argue that one of the advantages of working on machine ethics is that it might lead to breakthroughs in ethical theory since machines are well-suited for testing the results of consistently following a particular theory » (« Machinal Ethics : Creating an Ethical Intelligent Agent », *AI Magazine*, volume 28 Number 4, 2007, p. 15)

[16] La question du dilemme est une question classique de la philosophie morale. Le dilemme, c'est le problème du choix : que dois-je choisir ? Mais avant même la valeur morale d'un choix (bien ou mal), il y a le choix même comme condition de la morale. On passe de l'esthétique à l'éthique (ou morale) en passant de « et » à « ou ». Une vie esthétique, explique le philosophe danois Kierkegaard, se déroule sur le plan du « et » : la beauté n'est jamais exclusive. Je peux aimer sans contradiction une fugue de Bach et la cathédrale de Reims, et tout ce qu'on voudra d'autre : l'esthétique c'est le règne de : A et B et C et … Passer à la morale c'est, au contraire, dépasser la simple conjonction des préférences en se mesurant à une dimension nouvelle, l'alternative : ou bien A… ou bien B.

Si je choisis A, alors : (1) je fais passer A du possible au réel ; (2) je ne choisis pas B. (1) est la condition de la responsabilité — les conséquences de mon choix A me sont imputables, au moins dans les limites de ce que je savais et pouvais anticiper comme conséquences de A ; (2) est la condition de l'irréversibilité — à la limite, sans doute, je peux encore choisir B, mais il n'annulera pas le choix A et viendra après lui (B après et d'après A n'est pas identique à A ≠B)

Autrement dit, existentiellement le choix approfondit tragiquement la vie de l'individu : si je choisis A alors il est impossible que non A, maintenant et pour toujours. Ou il sera toujours vrai que A si j'ai choisi A. Ethiquement, l'existence se laisse décrire comme une suite ouverte de choix irréversibles (A puis B, puis C). L'existence de l'individu se laisse décrire comme l'histoire de ses choix en situation.

On peut donc dire, à ce premier niveau, que l'alternative (ou bien… ou bien) est la condition formelle ou générale de la morale et que l'alternative révèle le tragique de l'existence (responsabilité et irréversibilité).

Le dilemme moral est quelque chose de plus déterminé que l'alternative (comme condition formelle de la morale). Il soumet nos intuitions et nos théories éthiques à une sorte de test. On peut même dire que le dilemme est crucial en morale. Pour être consistante, une théorie morale doit-elle inclure ou exclure la possibilité du dilemme moral ? Si elle exclut le dilemme, elle paraît ne pas prendre en compte la réalité de la vie morale, de fait exposée à des conflits de devoirs ; si elle l'inclut, sans le résoudre, elle ne répond pas à sa fonction de fonder le choix moral. Donc si la théorie morale exclut le dilemme elle perd de vue la réalité de l'action : si elle l'inclut, elle risque de se nier comme théorie.

Mais on peut aussi, tout à l'inverse, relativiser l'importance du dilemme en morale, au point d'en nier l'existence. C'est d'ailleurs le lieu commun des théories morales, chez Platon pour l'Antiquité, Leibniz, Kant et Stuart Mill pour la philosophie moderne. Faut-il rendre une arme à son propriétaire s'il est devenu fou (Platon, *République* I, 331c) ? Rendre à son propriétaire l'objet qui lui appartient est un devoir. Défendre la sécurité en est un autre. Conflit des devoirs. Si le propriétaire n'était pas fou, l'obligation de rendre l'arme s'imposerait directement. Mais si celui-là est devenu fou, c'est le devoir de sécurité qui l'emporte. Dans ce cas le dilemme est soluble. Il suffit de hiérarchiser les devoirs selon un ordre de priorité (sécurité collective>propriété individuelle), ou de tenir compte des circonstances particulières (l'événement de la folie). Le dilemme entre les deux devoirs était un faux dilemme. Pour Thomas d'Aquin également, ce qui passe pour un dilemme moral résulte en réalité d'une mauvaise délibération de l'agent moral qui a mal ordonné les moyens pour la fin visée ou les devoirs relatifs entre eux (dilemme *secundum quid*). Et donc une théorie morale qui reconnaît l'existence de dilemmes en soi (*simpliciter*) est fausse.

On peut encore avancer un autre argument pour supprimer le dilemme. Il n'y a peut-être jamais égalité entre deux devoirs et donc jamais non plus aucun dilemme, faute d'identité entre les objets des devoirs ou du même devoir. Soit A et B en train de se noyer, avec A et B deux jumeaux. Apparemment le devoir de sauver A = le devoir de sauver B, puisque A = B. Mais en application du principe leibnizien de l'identité des indiscernables, il y a toujours une différence qui est la raison du choix. Si je sauve A plutôt que B de la noyade, c'est que j'avais une raison de le faire : par exemple, je suis plus proche de A, ou A est plus proche de la rive, ou A a encore la tête hors de l'eau, etc. Donc c'est un raisonnement rétrospectif mais illusoire qui me fait croire que j'avais le choix égal de A ou B ou le devoir égal de choisir A ou B.

Enfin, on peut considérer qu'à chaque fois qu'il y a dilemme, le critère moral n'a pas été appliqué correctement. Soit le cas de mon ami recherché par un homme qui veut le tuer et qui se réfugie chez moi. L'assassin frappe à ma porte et me demande si cet ami est dans ma maison. J'ai le choix entre ne pas mentir et livrer mon ami à son assassin ou sauver mon ami au prix d'un mensonge. Y a-t-il dilemme ? Non car soit je considère que la vie d'un homme (devoir 1) et qui se trouve être mon ami (devoir 2) prévaut sur tout, même sur le devoir de vérité ; soit, au contraire, comme le préconise Kant (*Sur un prétendu droit de mentir par humanité*, 1797), je considère que le devoir de vérité ou du moins de véracité est inconditionnel, contrairement à l'autre qui est conditionnel. Et même si je ne livre pas mon ami à son assassin, je sais que n'ai pas agi comme le devoir moral l'exige(r)ait. Donc pour Kant, il n'y a pas de conflit des devoirs parce qu'il n'y a qu'un seul devoir, l'impératif catégorique. Et si néanmoins je n'agis pas par devoir mais contre le devoir, je sais que j'ai fait un choix immoral :

toujours plus de nos choix, on peut craindre que l'éthique humaine ne finisse par imiter l'*IA* éthique. Mais surtout, même dans le cas du dilemme, le sujet peut, malgré un choix assumé comme acceptable, être frappé de scrupule, de remord, de honte. Il sait n'avoir pas agi comme le devoir lui commande d'agir. Il a conscience de n'être pas à la hauteur de la règle. L'*IA* pourrait progressivement effacer en l'homme la conscience de la culpabilité et de la faute. Ou alors elle l'en libérera. Sera-ce un bien ou sera-ce encore un homme ?

---

j'ai préféré ou fait passer des intérêts personnels avant le respect pour le devoir, j'ai lésé l'humanité (universel) en voulant sauver mon ami (particulier).

Le dilemme moral a retrouvé un certain crédit dans la philosophie morale contemporaine, autour du dilemme dit du tramway. Celui-ci a été imaginé par la philosophe britannique Philippa Foot en 1967. C'est une expérience de pensée, d'habitude présenté sous cette forme : une personne peut effectuer un geste qui bénéficiera à un groupe de personnes A, mais, ce faisant, nuira à une personne B ; dans ces circonstances, est-il moral pour la personne d'effectuer ce geste On peut considérer qu'un acte moral est rationnel s'il peut s'appuyer sur une raison, et si possible universalisable. Dans le cas du dilemme du tramway et de ses variantes, il paraît plus rationnel de sacrifier un pour sauver plusieurs. Ce critère est celui de la morale ou de l'éthique utilitariste. L'arbitrage est conforme à la maximisation de l'intérêt (de la fonction d'utilité) ou, ce qui revient au même, à la minimisation de la souffrance. Il est moral de réduire la quantité de souffrance et/ou d'augmenter la quantité d'utilité. La force de l'utilitarisme est de résoudre le problème du dilemme moral.

Mais le critère utilitariste est-il si convaincant ? On n'a pas manqué de critiquer l'expérience de pensée du tramway — il y a une littérature abondante en philosophie contemporaine, notamment anglo-saxonne, sur le dilemme, par exemple McConnell « Moral Dilemmas and Consistency in Ethics », *Moral Dilemmas*, New York, Oxford, Oxford University Press, 1987, et B. Williams, « Ethical Consistency » 1965 dans *La fortune morale*, Paris, Presses Universitaires de France, 1994.

1) sur un plan théorique, on peut contester la réduction de la morale à la psychologie, ou du moins l'abstraction de l'expérience de pensée, son caractère peu réaliste qui n'est d'aucun enseignement pour constituer un critère normatif de l'action ; 2) mais surtout certaines variantes du dilemme font apparaître des choses surprenantes, comme celle de l'obèse proposée par Judith Jarvis Thomson (*The Trolley Problem,* 1985) :

« Imaginez une situation — que j'appellerai « Homme obèse » — où vous êtes sur un pont sous lequel va passer un tramway hors de contrôle se dirigeant vers cinq ouvriers situés de l'autre côté du pont. Que faites-vous ? Étant un expert en tramways, vous savez qu'une manière sûre d'en arrêter un hors de contrôle est de placer un objet très lourd sur son chemin. Mais où en trouver un ? Au moment des événements, il y a un homme obèse, vraiment très obèse, à côté de vous sur le pont. Il est penché au-dessus du chemin pour regarder le tramway. Tout ce que vous avez à faire est de lui donner une petite poussée pour qu'il tombe sur les rails et bloque le tramway dans sa course. Devriez-vous donner cette poussée ? Tous ceux à qui j'ai posé cette question m'ont répondu non. Mais pourquoi ? ».

Mais le calcul utilitariste est-il vraiment moral ? Les variantes du dilemme font apparaître des conséquences surprenantes. 50% des individus interrogés qui approuvent le sacrifice de 1 individu pour 5, n'approuvent pourtant pas le geste dans cette présentation. Mais de manière encore plus inattendue, selon des études plus récentes menées par Joshua Green, professeur de psychologie de Harvard, si on propose de pousser l'innocent de sa passerelle non pas avec l'épaule ou la main mais avec une perche, beaucoup plus de personnes se déclarent prêtent à faire mourir l'inconnu — comme s'il était plus grave et moralement signifiant de pousser avec la main ou avec un bâton (cf. *Tribus morales*, Markus Haller, 2017). On peut tirer de ces variantes la conclusion que le sujet moral éprouve bien une tension entre des principes moraux (maximisation du bien, minimisation du mal *vs* interdit de tuer). Il semblerait ainsi que la morale emprunte deux circuits neuronaux différents : soit l'émotion, le mode « automatique », inconscient, peut-être sélectionné par l'évolution qui est la voie la plus courte, la plus générale, soit la raison critique, consciente qui calcule les conséquences, que Green nomme le mode « manuel ». Le raisonnement utilitariste n'est pas et ne peut passer ni en droit ni en fait pour le critère exclusif de l'action morale.

Pour ne pas terminer sur un doute et une tristesse, il y a une autre manière d'envisager le rapport entre l'*IA* et l'éthique, qui tourne le dos à notre hypothèse générale d'une transformation de l'éthique par l'*IA* et qui rappelle (avec B. Williams) que l'éthique ne se réduit pas aux concepts minces des théories normatives (devoir…). L'éthique c'est aussi des expériences, des relations faites d'affects, de dispositions, d'empathie, de souci, de bienveillance qui ne relèvent ni de l'impératif (morale déontologique), ni du calcul (utilitarisme), ni (de la forme générale) du dilemme. On travaille actuellement beaucoup à la production d'une « empathie artificielle », permettant à l'*IA* à réagir de manière appropriée à l'humeur d'une personne, propice à instaurer une relation plus conviviale entre l'homme et la machine.

Mais une relation plus éthique entre l'humain et la machine ne suffit pas à faire de la machine un agent moral authentique. Ou plutôt la relation éthique suppose la relation entre deux sujets ou deux agents moraux. Est-ce le cas entre l'humain et l'*IA*, même doté d'un comportement éthique ? Une machine éthique ou un *APAM* est-il un agent moral comme l'humain avec lequel une relation éthique doit être instaurée[17] — relation à double sens, du côté de l'éthique artificielle (*IA*) par la programmation éthique de la machine ; du côté de les agents moraux humains pour apprendre à vivre et à interagir avec des nouveaux agents pratiques ?

Le robot (*APAM*) peut avoir une apparence humaine, simuler la douceur, stimuler intellectuellement et affectivement des patients, faire du bien (soulager, soigner, rassurer), en lisant sur le visage la détresse, la tristesse, en répondant à l'urgence médicale, etc., il ne peut être considéré comme un « agent moral artificiel »[18]. Un *APAM* est certes un « agent pratique artificiel » : il agit sur les choses et en agissant il fait le bien ou du bien aux humains. Mais c'est sans le savoir, et s'il fait le mal, c'est sans le vouloir. Il n'a pas plus de motivation que de plaisir à bien faire et il ne juge pas ses actions. Autrement dit, c'est à la fois un « zombie sans émotions ni passions » (ou un « saint apathique ») et un « ingénu ». On doit sans doute lui attribuer une responsabilité pratique propre, dès lors qu'il ne se contente pas (plus) d'exécuter comme un automate un acte programmé par son concepteur, mais qu'il déploie une capacité d'initiative et donc une imprévisibilité dans des situations typiques de choix. Cette responsabilité pratique n'est pourtant pas encore morale, tant il manque à l'*APAM*, de pouvoir accéder aux raisons, aux justifications et aux excuses des actions et des effets des actions dont il est la cause. Faute de disposer de la « profondeur sémantique » pour saisir la différence simple mais décisive entre *is* et *ought*, des capacités affectives qui la rendent possibles, il restera comme un éternel enfant, un agent pratique artificiel amoral, c'est-à-dire un agent pratique incapable de se perfectionner moralement. Nos relations éthiques avec ces agents pratiques artificiels non moraux exigeront de nous plutôt l'indulgence, la patience que le blâme[19].

[17] La relation éthique est à double sens : l'IA doit programmer la machine d'un comportement éthique intelligent, pour s'adapter au monde éthique humain ; les agents moraux humains doivent apprendre à vivre et à interagir avec ces nouveaux agents pratiques.

[18] Nous suivons à nouveau les analyses de Stéphane Chauvier, *art. cit.*, fin.

[19] S. Chauvier conclut ainsi son article : « Vivre avec des APAM, c'est vivre entouré de petits agents pratiques bienveillants, mais parfois maladroits. Pour considérer sans superstition ces nouveaux habitants du monde, il nous faudra sans doute apprendre à ne pas leur prêter une dignité qu'ils n'ont pas, même s'ils sont dotés d'une apparence humaine : ils n'auront aucun mérite à nous faire du bien, nous n'aurons aucune gratitude ou reconnaissance à avoir

Des agents d'un nouveau type, artificiels et intelligents, vont peupler notre monde et intervenir dans et entre nos vies. Comment les humains, jusqu'à présents les seuls agents moraux au monde, vont-ils devoir apprendre à se comporter avec eux ? Cette question pragmatique suppose elle-même de savoir si un agent pratique artificiel peut et doit être considéré comme un agent moral de plein droit. Et cette question suppose à son tour de savoir déterminer quel type d'entité est par exemple, un véhicule autonome, ou un robot autonome aide-soignant. Ainsi la question pratique : « comment vivre avec les *IA* et les robots autonomes ? », implique, on l'a vu, la question morale : « l'agent pratique artificiel est-il un agent moral ? » — ou y a-t-il un agent moral artificiel possible ? —, qui, à son tour, pose une question ontologique inédite : « quel type d'entité est une *IA* ? »

Bibliographie

1965 : Bernard Williams, « Ethical Consistency », *La fortune morale*, Paris, Presses Universitaires de France, 1994.

1967 : Philippa Foot, « The Problem of Abortion and the Doctrine of the Double Effect », *Virtues and Vices*, Oxford, Basil Blackwell, 1978

1985 : Judith Jarvis Thomson, *The Trolley Problem*

1987 : McConnell, « Moral Dilemmas and Consistency in Ethics» , *Moral Dilemmas*, New York, Oxford, Oxford University Press,

2007 : D. Levy, *Love and Sex with Robots*, New York, Harper Collins, 2007.

2007 : Michael et Susan Anderson, « Machinal Ethics : Creating an Ethical Intelligent Agent », *AI Magazine*, volume 28 Number 4, 2007

2014 : «Algorithm Aversion : People Erroneously Avoid Algorithms After Seeing Them Err », *Journal of Experimental Psychology*

2016 : Stéphane Chauvier, « L'éthique artificielle », *l'Encyclopédie philosophique* : http://encyclo-philo.fr/ethique-artificielle-a/

2017 : Joshua Green, *Tribus morales*, Markus Haller

2017 : CNIL, « Les enjeux éthiques des algorithmes et de l'intelligence artificielle »

2017 : Will Knight, « The Dark Secret at the Heart of AI », *MIT Technology Review*

2019 Martin Gibert, « L'éthique artificielle », *l'Encyclopédie philosophique* : http://encyclo-philo.fr/etique-artificielle-gp/

envers eux : nous les aurons faits pour ça. Et nous ne pourrons jamais, nous-mêmes, leur faire du mal, parce qu'ils seront des zombies ingénus. En revanche, nous devrons être prêts à faire montre d'indulgence à l'égard de leurs bévues. Tandis que nous pouvons remiser sans vergogne un toasteur qui brûle systématiquement nos toasts, nous devrons faire montre d'un peu de patience à l'égard de nos APAM domestiques : ils devront apprendre à vivre parmi nous, apprendre à nous connaître, et ils pourront donc commettre toutes sortes d'erreurs avant de mettre leurs actes en harmonie avec leur volonté sainte. Par prudence cependant, il nous faudra contracter une assurance en leur nom, car les dommages qu'ils causeront devront être couverts. Mais le règne des fins leur restera pour longtemps fermé ».