**Colloque NXU, 5 juin 2018, TBS**
**Laurent Cournarie, « L'éthique (d') après l'IA »**

## Introduction

Bonsoir,

Le thème « IA et éthique » commence à être très débattu. Il voit s'opposer deux positions. Certains peuvent considérer que la réflexion éthique en IA est un thème marketing 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. D'autres au contraire considérant que certaines valeurs éthiques sont essentielles à l'humanité, s'élèvent contre la naïveté à croire que les algorithmes sont neutres, qu'ils ne recèlent pas de biais cognitifs, que les résultats en *deep learning* sont transparents et explicables. La réflexion éthique se présente alors comme une intelligence critique de l'IA.
Le débat sur l'éthique et l'IA porte en général sur les effets éthiques de l'IA. On en pointe 4 principalement :
- nécessité de protéger les données personnelles parce que l'éthique commence par le respect de la liberté individuelle
- nécessité de lutter contre les discriminations parce que l'éthique repose sur le principe de dignité et d'égalité des personnes[1]
- nécessité de préserver le pluralisme humain contre une tendance à "clusterisation" des individus parce l'éthique reconnaît à l'individu de pouvoir être le sujet pour soi de sa vie propre
- nécessité de défendre le bien commun ou l'intérêt général des sociétés et même de l'humanité en général à la fois contre les entreprises privées et contre l'individualisme de l'usager et du consommateur lui-même. Autrement dit, l'éthique est toujours une question politique.

Mais plus précisément, si l'on ne confond pas l'éthique avec ce qui n'est pas technique et avec le sociétal, qu'en est-il de l'éthique d'après l'IA, càd à la fois que peut être l'éthique en IA et que peut faire l'IA à l'éthique ?
Dans cette communication, je propose 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 est celui de l'autonomie humaine de la décision

[1] 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 peu claire était un critère de beauté.

## (1) Un monde plus éthique.

Le monde est plus éthique 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 le monde est plus éthique parce que le périmètre de l'éthique s'est indéfiniment élargi. Qu'entend-on par éthique ? Une théorie normative de l'action. Les hommes et les sociétés ont toujours considérés que toutes les actions ne se valent pas : ils 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 bien et du mal, du bon et du mauvais. Mais cette aptitude à norme l'action jusque-là avait une extension limitée. L'éthique concernait :

1. le rapport de l'homme à l'autre homme — étaient exclus de l'éthique la nature, l'animal, la machine : ce n'est plus le cas.
2. une temporalité resserrée autour du présent de l'action (présent du rapport à autrui, responsabilité de l'action passée ou de la conséquence prévisible — étaient exclus de l'éthique les générations à venir et les effets imprévisibles de la puissance humaine sur la nature : ce n'est plus le cas.
3. les limites de la finitude humaine — étaient exclus de l'éthique tout ce qui touche aux fins de l'existence (naissance/mort) : ce n'est plus le cas[2].

Le monde est plus éthique parce que l'humanité est plus responsable : l'humanité est responsable de ce qui n'avait jamais relevé de l'éthique. *From chance to choice*, comme on répète.

## (2) Un monde moins moral

Mais en même temps le monde est peut-être moins moral, au sens où on assiste peut-être 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'ancien monde était fait de devoirs et d'interdits, investis et cautionnés par des institutions. Or pour des raisons multiples[3], il est de moins en moins possible de définir une règle universelle et/ou de l'imposer durablement. Comme si le "non" s'effaçait tendanciellement de notre monde. Les intérêts économiques, la disparité des systèmes juridiques, le désir d'apprendre et de connaître davantage sont plus puissants que tous les scrupules moraux.

Ces remarques très générales sur l'éthique étant faites, venons-en à la 3ème idée sur l'éthique (d') après l'IA.

## (3) Autonomie humaine de la décision.

[2] Descartes écrivait dans une lettre : « 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 » (Lettre à Chanut,15 juin 1646). Maîtriser ses désirs, ses craintes, acquérir des vertus, voilà ce que l'éthique exigeait de l'homme, faute de pouvoir maîtriser le processus de sa vie biologique.

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

A mon sens, l'enjeu éthique de l'IA se concentre sur l'autonomie humaine de la décision [4]
Une bonne illustration de cet enjeu est le cas de l'automobile autonome. Le problème éthique de l'automobile autonome (que je propose de nommer l'autonomobile) concerne la programmation d'une machine à choisir en situation critique : comment l'homme doit-il programmer la machine qui le dispensera à l'avenir de choisir ?
Le problème éthique est en fait double : d'abord l'individu est dépossédé de sa faculté de décision. Mais le plus singulier c'est qu'on programme une machine à un choix 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'accident se présente comme un événement, que chacun négocie comme il peut, dans des circonstances contingentes (sa personnalité, son aptitude à la conduite, 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 : l'IA doit être programmée à choisir A ou B, et on peut concevoir une "autonomobile" ou exclusivement altruiste (choisissant de sauver systématiquement les piétons par exemple, même au risque de la mort des passagers) ou exclusivement égoïste (choisissant de sauver systématiquement les passagers au risque de la mort des piétons).
Et pour la programmer, on accumule les données à partir des réponses à plusieurs scénarios pré-définis comme sur la plateforme du MIT, *Moral machine*, en fonction de variantes : l'âge (enfants ou personnes âgés), la nature (personnes ou animaux), la position (passagers ou passants), le nombre (qui combine les critères précédents).

Le cas de la voiture autonome appelle plusieurs remarques plus générales sur l'IA et l'éthique :

(1) une considération "humaniste" consolatrice : il n'y a pas de problème éthique pour l'IA : le dilemme n'a pas de sens moral pour l'IA. A moins d'une IA forte, càd d'une conscience de soi artificielle et donc (c'est moi qui l'ajoute) d'une conscience morale artificielle, le choix de l'IA n'a de sens moral que pour l'homme. L'IA ne choisit pas et n'agit pas éthiquement mais suit un algorithme programmé selon des normes éthiques pré-définies par l'homme et signifiantes seulement pour lui. Mais que se passera-t-il si l'on parvient à développer complètement un « agent éthique explicite » artificiel et non plus seulement implicite (suivant une procédure rigide), capable de calculer l'action la meilleure en situation critique de dilemme moral selon des principes moraux ?
(2) une interrogation : que se passe-t-il si, malgré tout, l'IA est défaillante ? On ne peut parler de faute morale puisque l'IA n'agit pas moralement : on a affaire à une erreur, une panne, un bug. Autrement dit, dans le monde qui vient, la valeur technique (fonctionnement/dysfonctionnement) risque de remplacer toujours davantage la valeur morale de l'action (bien/mal).
(3) une inquiétude : on peut se demander si on ne tend pas à substituer un dispositif technique (IA) à une responsabilité éthique (humaine). Imaginons un parc automobile intégralement composé de véhicules autonomes : il n'y aurait plus (ou quasiment plus) d'accidents. Ce qui tendrait à prouver que l'exigence de responsabilité éthique n'est que l'envers de la faillibilité humaine : et si on résout le problème de la faillibilité humaine (par la voiture autonome), on résout le problème de la responsabilité et du choix éthiques, mais en le faisant disparaître. En bref, l'IA pourrait à terme dispenser de l'éthique.

[4] Le rapport de la CNIL se demander « comment permettre à l'homme de garder la main ? ».

## Conclusion

Alors que peut faire l'IA à l'éthique ? Je me contenterai de souligner 4 points pour terminer :

1. une remarque générale : l'éthique change avec la puissance de la technoscience (c'est la thèse du philosophe allemand Hans Jonas). On doit peut-être remettre en cause l'idée, admise largement, que la technique serait par définition éthiquement neutre (technique = moyen/éthique = fin). Le dispositif technique dans le cas de la technoscience et de l'IA induit des modes de penser et d'évaluer qui ne sont pas neutres.
2. une question : quelle éthique est programmable dans une IA ? En IA, l'éthique prend nécessairement la forme du dilemme (popularisé par l'expérience de pensée de Philippa Foot) et le dilemme est posé en termes d'acceptabilité[5]. Or l'acceptabilité n'est pas le critère éthique en soi mais le critère de l'éthique utilitariste : est moral ce qui est acceptable ce dont la conséquence constitue la plus grande somme de biens ou la moins grande somme de maux. Or l'utilitarisme est une théorie éthique (normative) et non pas toute l'éthique. Donc l'IA rencontre l'éthique principalement en sélectionnant une théorie normative — mais pour une raison évidente : elle est la seule qui se prête à la quantification (+ de maux/- de maux). Impossible d'introduire dans un algorithme un commandement inconditionnel : "tu ne tueras point" ou la règle d'Or qui est sans doute la règle éthique la plus universelle (« ne fais pas à autrui ce que tu ne voudrais pas qu'on te fasse ».
3. comment interpréter l'éthique en IA ? On le peut de deux manières sans doute :
(a) selon une interprétation positive, l'IA est l'occasion de clarifier nos intuitions éthiques. La recherche en IA et en IA éthique constitue un progrès en éthique[6].
(b) selon une interprétation plus prudente, on peut craindre sinon un appauvrissement de l'éthique dans toute son expression humaine, du moins une éthique diminuée. Car même si une IA peut choisir comme aurait choisi un agent humain, l'agent humain peut en éprouver du remord, de la culpabilité. L'éthique d'après l'IA est une éthique privée de la conscience de la faute. Or la conscience de la faute constitue peut-être le foyer même de l'éthique.
4) Mais pour ne pas terminer sur un doute et une tristesse, on peut envisager le rapport entre l'IA et l'éthique, d'une manière plus positive. L'éthique est faite de relations entre les êtres, de dispositions affectives. Or d'un côté, l'IA travaille 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. De l'autre, les agents humains vont sans doute devoir apprendre à vivre avec l'IA et à développer avec elle des relations éthiques. Donc l'éthique dans un monde IA est possible.
Mais la relation entre un agent éthique de plein droit (humain) et ce qu'un philosophe français contemporain, Stéphane Chauvier, nomme un « agent pratique artificiel modulaire »

[5] Cf. *Moral Machine* et M. et S. Anderson, « Machinal Ethics : Creating an Ethical Intelligent Agent », *AI Magazine*, volume 28 Number 4, 2007

[6] 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 pachine ethics is that it might lead to breakthroughs in ethical theoryn since machines are well-suited for testing the resultas of consistently following a particular theory » (« Machinal Ethics : Creating an Ethical Intelligent Agent », *AI Magazine*, volume 28 Number 4, 2007, p. 15)

(*APAM*[7]) est-elle pleinement éthique ? Car par comparaison avec l'humain, l'IA (*APAM*) est un zombie sans émotions, ni passions (c'est un saint apathique, sans mérite, qui fait du bien sans la tentation du mal) et un agent ingénu, incapable de juger, d'examiner son action. Mais l'APAM sera-t-il toujours cet agent éthique déficient, c'est toute la question.

Merci de votre attention.

---

[7] Une IA est un agent pratique artificiel modulaire (*APAM*) : /1/ agent pratique, parce qu'il poursuit de buts pratiques intelligemment ; /2/ agent artificiel, parce cette intelligence pratique est programmée informatiquement ; /3/ agent modulaire parce qu'il ne couvre par rapport à un agent pratique humain qu'un segment limité de buts pratiques (cf. Stéphane Chauvier, « Ethique artificielle », *Encyclopédie philosophique*).