Quelle éthique pour l'IA ?

Thierry Ménissier

▶ **To cite this version:**

Thierry Ménissier. Quelle éthique pour l'IA ?. Naissance et développements de l'intelligence artificielle à Grenoble, Académie Delphinale, Oct 2019, Grenoble, France. halshs-02398215

HAL Id: halshs-02398215
https://halshs.archives-ouvertes.fr/halshs-02398215

Submitted on 7 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quelle éthique pour l'IA ?

Colloque de l'Académie Delphinale

Grenoble, 19/10/2019

Thierry Ménissier

Professeur des Universités, philosophie
Responsable de la chaire « éthique & IA »,
Multidisciplinary Institute in Artificial Intelligence
https://miai.univ-grenoble-alpes.fr/

Institut de Philosophie de Grenoble
https://iphig.univ-grenoble-alpes.fr/
Université Grenoble Alpes
Bâtiment ARSH, BP 47, Domaine Universitaire, 1281, avenue Centrale
38400 Saint-Martin d'Hères
Thierry.menissier@univ-grenoble-alpes.fr

> « Le monde était si récent que beaucoup de choses n'avaient pas encore de nom et pour les mentionner, il fallait les montrer du doigt. »
>
> Gabriel Garcia Marquez,
> *Cent ans de solitude*,
> trad. Claude et Carmen Durand,
> Paris, Editions du Seuil, 1968.

Nous entendons souvent par « intelligence artificielle » l'activité des algorithmes dont les opérations de calcul sont nourries par des flux de méga-données (*big data*) eux-mêmes engendrés par des capteurs variés. Ces technologies de calcul se complètent de deux prolongements qui engagent des éléments a priori étrangers à elles, prolongements qu'il est souvent difficile de distinguer ce qui relèverait de l'IA conçue en un sens restreint, c'est-à-dire seulement le travail algorithmique, et en un sens large, c'est-à-dire l'IA complétée : elles se fondent d'un côté sur les technologies de l'information et de la communication et appuient l'essor du numérique, et de l'autre elles s'expriment sous les diverses formes actuelles de la robotique. Ce système technique qu'aujourd'hui nous appelons IA, nous pourrions simplement le nommer « informatique avancée » ou « informatique augmentée par les données » ; ce n'est pas encore une véritable « intelligence artificielle » comme il existe des intelligences naturelles, c'est-à-dire des entités capables de prendre des décisions de manière autonome en fonction d'une réflexion consciente.

Sur un sujet technologique comme celui de l'IA ainsi définie, il peut sembler incongru d'inviter un philosophe à parler (plutôt qu'un sociologue par exemple), et plus encore de l'inviter à parler en dernier, comme si, pour clore les débats, il fallait convoquer une parole pleine de « sagesse », la philosophie étant la recherche de la *sophia*, et, puisqu'on la convoque à la fin du colloque, une parole capable d'avoir le dernier mot. Et de fait, sur bien des sujets contemporains où s'expriment des formes d'inquiétude vis-à-vis de l'avenir, il existe aujourd'hui une sorte de tentation : la parole philosophique paraît interpellée dans un rôle particulièrement difficile à tenir pour elle (la chouette, l'animal d'Athéna, se levant au crépuscule, comme le rappelait Hegel), celui de clore tous les débats relatifs aux technologies

nouvelles qui engagent un futur non seulement par nature inconnu, mais dont on dit de plus que, du fait des nouvelles technologies, il sera différent de tout ce qu'on a connu par le passé.

D'un autre côté, lorsqu'on veut associer les termes « éthique » et « IA », il apparaît en effet impossible de ne pas convoquer la philosophie, tant la situation semble complexe, confuse et paradoxale. Si par rapport aux autres disciplines académiques, la philosophie n'a pas le monopole de la complexité, à l'instar des autres elle vise à dissiper la confusion sur les sujets qu'elle aborde (en clarifiant les concepts, en formulant des problèmes et en proposant des argumentations cohérentes), et elle présente l'avantage (sans doute est-ce dû à son fondateur Socrate, un Athénien ironique) d'être plus que tout autre disposée à assumer les paradoxes, ces tensions indépassables entre des thèses à la fois essentielles et irréductiblement contradictoires les unes avec les autres. Or, concernant notre situation actuelle face à l'IA, la confusion apparaît réelle, et on pressent qu'elle peut nourrir certains paradoxes. Dans cette contribution, je veux surtout m'attacher à dissiper la première en clarifiant successivement trois points. J'aborderai successivement les relations entre éthique et IA dans l'expression « éthique de l'IA », puis le besoin d'éthique vis-à-vis des situations contemporaines où des solutions d'IA sont déployées, enfin le type d'éthique propre à satisfaire ces besoins. Ce cheminement me conduira à souligner la nécessité de se positionner sur un autre plan, où on attend également quelque chose du philosophe sur les sujets, même technologiques, qui engagent l'avenir : celui de la réflexion sur les enjeux, les valeurs et les finalités poursuivies par l'action humaine.

1. *Clarifier la confusion de relations entre éthique et IA : que signifie « éthique de l'IA » ?*

Aujourd'hui, de nombreux acteurs de la société en expriment le besoin ou affirment qu'il y a un besoin urgent ou qu'il est nécessaire de réfléchir éthiquement à l'IA, voire de constituer une « éthique de (ou pour) l'IA » : des citoyens [Montréal 2018], des ONG [Toronto 2018], des institutions politiques [CNIL, 2017 ; European Commission 2019], et même des sociétés productrices d'IA [Google 2018 ; Microsoft 218] en ont récemment exprimé le besoin. Mais la notion même d'une « éthique de l'IA » est tout sauf évidente. Car comment entendre ce « de » ? Rigoureusement parlant, l'orientation même du questionnement change en fonction de l'acception de la particule dans l'expression « éthique de l'IA ». Et de fait nous pouvons évoquer deux acceptations possibles de la particule, qui engagent des types assez différents de réflexion.

Premièrement, l'éthique *appliquée à* l'IA ou éthique *pour* l'IA dont on parle relève-t-elle de l'éthique générale, ou bien est-elle spécifique ? Si elle relève de l'éthique générale, alors nous serions aujourd'hui dans le cas, bien connu, de ce qui se produit quand apparaît une nouvelle technologie, par exemple lorsque l'électricité a commencé à révéler son potentiel, au XIX^{e} siècle, ce qui a engendré des émotions fortes et contradictoires, jusqu'à créer une nouvelle mythologie, ou du moins une sorte d'amendement ou de révision de la mythologie héritée[1]. Dans ce cas, le contexte se traduit à la fois par une tout à fait normale incertitude (car on ignore les applications possibles d'une découverte ou d'une invention prometteuses mais trop récentes pour être maîtrisée) et par une non moins légitime inquiétude (personne en effet ne peut dire ce que l'innovation qui surgit va modifier dans les usages ou les mœurs). Dans un tel contexte, toutefois, les questions supposées éthiques qui viennent à être formulées

1 Cas intéressant, le récit *Frankenstein* de Mary Shelley [Shelley, 1818] qui évoque une créature artificiellement constituée de chair inerte et d'organes empruntés à des défunts ranimés, se présente comme le symptôme d'une telle émotion qui s'est cristallisée avec l'apparition de l'intérêt scientifique pour l'électricité, et il a au fil de ses déclinaisons cinématographique et romanesque dans ce mouvement émotionnel reformulé la mythologie héritée des Grecs (l'histoire de Prométhée) et des Juifs (la créature du Ghetto de Prague appelée le Golem).

concernent seulement ce qu'on pourrait nommer le « réglage » des usages sociaux et des mœurs qui se trouvent modifiés (cela peut être en profondeur ou superficiellement) par le déploiement de la technologie. À cet égard, il convient de remarquer, en convoquant l'histoire des techniques, que l'apparition de certaines technologies à la fois révolutionnaires, fondatrices de nos sociétés et objectivement meurtrières n'a pas suscité un tel émoi ; on peut penser ici à l'apparition puis à la diffusion rapide du moteur thermique fonctionnant à l'énergie fossile non renouvelable et à sa déclinaison massive dans les transports routiers, notamment sous forme de véhicule privatifs.

Mais, deuxièmement, si l'éthique de l'IA est spécifique, jusqu'où l'est-elle ? Peut-on imaginer que l'éthique de l'IA soit entendue comme l'ensemble des règles partagées par une corporation professionnelle, celle des informaticiens, ou l'ensemble des questions non directement informatiques examinées et débattues par les informaticiens, à l'instar de la bioéthique pour les médecins ? L'éthique de l'IA pourrait alors être perçue comme une sous-branche de l'informatique, de même que la bioéthique peut, à certains égards, être considérée comme une branche de la médecine. Ce qu'on appelle la *Computer Ethics* serait, rapportée à l'IA, l'équivalent de la bioéthique rapportée à la médecine. On peut souligner deux aspects très intéressants de cette manière d'envisager les choses, (1) elle indique que dans les sujet « éthique et IA », il y a une forte dimension scientifique et technique : si l'on veut réellement entendre quelque chose en bioéthique, cela implique une certaine connaissance médicale, pareillement en éthique de l'IA, il convient de se familiariser avec l'état de l'art informatique ; (2) ce qu'on pourrait nommer l'attraction de l'éthique par les informaticiens spécialistes d'IA engendre elle-même une conséquence. Le médecin peut ici donner quelques conseils à l'informaticien, car il a en quelque sorte une longueur d'avance sur lui : cette nouvelle situation induit pour les gens de l'art un accroissement de leur responsabilité, car ce sont eux qui devront, face aux usagers, rendre raison des règles et du sens général qui aura été donné à tel ou tel développement technique.

Les deux acceptions de l'expression « éthique de l'IA » que je viens d'évoquer renvoient-elles à ce qu'on définit généralement comme éthique, à savoir, le type de pensée visant à établir des relations entre le sens qu'un humain attribue à ses décisions ou à ses actions et les règles qui découlent de ce sens ? C'est que ce je vais examiner à présent.

2. *Clarifier le besoin d'éthique vis-à-vis des développements de l'IA : la remise en question des relations entre agentivité et responsabilité*

Il apparaît que, même si l'IA d'aujourd'hui n'était qu'une « informatique avancée » (ou « IA faible »), les développements technologiques contemporains posent en effet des problèmes éthiques.

Si l'éthique est requise, c'est que le déploiement de l'IA a d'ores et déjà mis en œuvre des pratiques où tant la capacité humaine de décision et d'action (ce que les philosophes anglo-saxons nomment l'agentivité[2] [Schlosser 2015]) que la responsabilité (dans le sens non juridique que retient Ricœur [Ricœur 1995]) sont susceptibles de se transformer de manière inquiétante. L'inquiétude vient de fait que les déploiements techniques nous font craindre d'être à la fois *moins libres* et *moins responsables* qu'autrefois. Le point crucial est que ce qui paraît aujourd'hui devoir être remis en question, c'est la subtile partition qui avait lentement été établie au cours des siècles passés, tout au long de la construction philosophique de la société moderne, entre ce que l'humain projette librement de penser et de faire et ce dont il doit répondre devant les autres.

[2] L'agentivité (*agency*) est un terme philosophique non connoté moralement qui désigne la capacité individuelle de s'identifier soi-même comme l'agent d'une action ; elle qualifie un être dont il est, par principe, toujours possible qu'il soit reconnu, au moins dans ses intentions, auteur de son acte.

De fait, si les ordinateurs font depuis plus de cinquante ans partie des outils qui appuient l'activité humaine dans de très nombreux domaines (et les calculs algorithmiques depuis bien plus longtemps), les nouveaux systèmes d'apprentissage (*deep learning*), la mise en système des algorithmes et l'apparition de « systèmes d'IA » (SIA) produisent aujourd'hui des effets de seuil tangibles en termes d'automatisation de l'activité, de prise en charge de celle-ci par la machine, ce qui se traduit par des nouvelles situations vécues dans de très nombreux secteurs importants. Si, pour les activités concernées, l'apport massif de solutions d'IA est présenté comme un surcroît considérable d'efficacité, ces situations peuvent parfois apparaître comme inouïes, comme radicalement originales pour les personnes engagées dans l'activité. D'une part, elles leur font perdre de vue les points de repère traditionnels sur ce qui est bon, juste et bien, de sorte que la réflexion éthique et politique se trouve sollicitée ; de l'autre, la prise en charge de l'activité par les machine modifie le schéma classique de l'imputation qui permet d'assigner les responsabilités. Une revue, même rapide, des transformations actuellement en cours dans les usages, fait bien apparaître que des questionnements de type philosophique (engageant la réflexion tant sur la liberté ou plutôt l'agentivité que sur la responsabilité) sont aujourd'hui appelés par tous les secteurs.

Dans le secteur de la *mobilité,* les développements technologiques en cours des véhicules autonomes remettent en question la maîtrise directe des engins par la volonté humaine. Ils engendrent une effervescence déjà traduite par une abondante littérature, produite en vue d'assurer une clarification des nouvelles conditions d'exercice de l'agentivité-responsabilité en régime de mobilité intégralement ou partiellement automatique [cf. par exemple Lin 2015 ; Perrin 2019]. Dans celui des *soins* pour les personnes en difficulté, « l'internet des objets » (*Internet of Things*) modifie la partition établie entre vulnérabilité et autonomie, en conférant aux SIA le pouvoir de régir de manière éventuellement très soutenue la vie des personnes en difficulté. Une véritable injonction est alors faite aux soignants comme à l'entourage familial des personnes protégées, afin qu'elles déterminent quelle limite assigner à leur responsabilité compte tenu de l'efficience incontestable de cette nouvelle domotique : de la surveillance à l'assistance corporelle, il est aujourd'hui techniquement possible que les machines gèrent intégralement l'existence des personnes en déficience d'autonomie, les solutions technologiques déjà en vigueur permettant de mettre en œuvre pour ces dernières une agentivité de substitution. Concernant la *défense*, l'apparition de machines de guerre autonomes, qu'il s'agisse des drones pilotés à distance du terrain d'action [Chamayou 2013] ou de « robots-tueurs » [Germain 2019], bouleverse l'art de la guerre. Les performances de l'IA conjuguées à l'essor de la robotique, dans le contexte du déploiement de stratégies militaires soucieuses d'économiser la vie des personnes humaines engagées sur le terrain, conduisent probablement ce secteur à devoir subir des transformations radicales, à partir de la promesse délivrée par l'apparition de « substituts » robotiques, ontologiquement définis comme des êtres intermédiaires entre les objets techniques et les acteurs sociaux [Dumouchel & Damiano 2016 : 31-68]. Enfin, dans celui de la *sûreté*, il est évident depuis quelque temps que les expériences de reconnaissance faciale engagent des formules technologiques de vidéosurveillance intelligentes qui obligent les pouvoirs publics à redéfinir les périmètres des sphères publiques et privées [Brey 2004]. Actuellement, l'évolution de la surveillance biométrique en régime de méga-données pose des problèmes d'une telle complexité concernant la qualification même de la notion de contrôle qu'ils impliquent a minima un nouveau pacte d'association entre sciences technologiques et sciences sociales [Marciano 2019]. Faute de quoi, et à l'instar du domaine de la défense, ce domaine voit émerger des formes d'innovation susceptibles de produire une substitution entre l'humain et la machine.

Dans le secteur de la *santé,* le développement de la médecine personnalisée, préventive et prédictive, c'est-à-dire du savoir médical combinant les connaissances génétiques et la puissance de prédiction du calcul, semble devoir en quelque sorte ajouter un pan entier à la

bioéthique dans lequel il s'agit, pour chaque sujet vivant, d'évaluer des situations inouïes pour sa responsabilité en fonction de l'actualisation potentielle de son propre patrimoine génétique hérité [Guchet 2016]. Dans ce cas, l'agentivité-responsabilité humaine se trouve en quelque sorte abyssalement étendue aux possibles (par exemple, des parents peuvent être tenus responsables de ne pas avoir traité au niveau du fœtus une déficience potentielle de la santé de leur futur enfant, qui aurait été attestée dès les premières analyses génétiques subie par ce dernier). Concernant ceux de *la production et de la gestion de l'énergie,* la diffusion des réseaux de distribution d'électricité qui favorisent la circulation d'information entre les producteurs, les fournisseurs et les consommateurs afin d'ajuster le flux d'énergie en temps réel et permettre une gestion plus efficace de l'offre et de la demande (*smart grid*), est de nature à bouleverser non seulement les rapports de pouvoir qui structurent les territoires, mais également elle contribue à renouveler la problématique de la souveraineté[3]. Conjugués à l'essor des énergies renouvelables, les systèmes énergétiques intelligents transforment notamment les rapports entre citoyens et pouvoirs publics, désamorçant les fonctions régaliennes autrefois dévolues aux régies nationales dans la production, la distribution et la gestion de l'énergie. Ce qui se profile ici, c'est la fin des standards étatiquement régulés de production et d'acheminement de l'énergie, et par suite émerge un questionnement sur la responsabilité des particuliers équipés de capteurs (voltaïques, éoliens, géothermiques) et de moyens alternatifs de production. À l'échelle d'un quartier ou d'une commune (soit aux niveaux où se joueront désormais des questions autrefois d'ordre régalien), ces individus peuvent devenir pourvoyeurs d'énergie pour une communauté : aussi, quels modes adopter pour la diffusion de l'énergie qu'ils ne consomment pas eux-mêmes ? Quelles règles et quels critères de répartition pour quelle forme d'équité ?

En matière de *transactions économiques et financières,* la *blockchain* (conçue indépendamment des monnaies virtuelles qu'elle permet de mettre en circulation) constitue une « chaine de blocs » capable de certifier les transactions à partir de postes informatiques indépendants les uns des autres et anonymes les uns vis-à-vis des autres. Elle promet de remplacer les tiers de confiance qui, depuis les premiers débuts des échanges financiers de grande ampleur, avaient pour mission de sécuriser les échanges, tels que les notaires ou les banques centrales. Décentralisée et dématérialisée, la chaine de certification des transactions valide celle-ci. L'ambigüité réside dans le fait qu'en l'état du développement du ce genre de système, ce qui accrédite la chaine, c'est à la fois la sûreté des machines qui codent et le travail de chaque « mineur », maillon-opérateur de base des transactions [Mallard *et alii* 2014 ; Caseau & Soudoplatoff 2016]. Parce que la « confiance distribuée » que génère la *blockchain* repose à la fois sur une forme de déterminisme technique et sur une reformulation, certes, mais également sur une réaffirmation du rôle d'humains garants de la transaction, la chaine de consensus qui se reconstruit sous l'effet de cette innovation comprend pour l'instant une indétermination majeure en matière de reformulation de l'agentivité-responsabilité.

En urbanisme et aménagement, l'essor des « *smart cities* » démultiplie et rend irréductible l'ambigüité que nous venons de mentionner. En effet, non seulement la variété des modèles en cours de développement sous cet unique terme produit une sorte de confusion, mais la réflexion sur nouvelles formes d'habiter et d'investir l'espace urbain, provoque un débat de fond sur la cité comme lieu traditionnel de socialisation et de politisation [Picon 2013]. Bien entendu, malgré son efficience en matière de gestion, la « ville intelligente » ne peut être uniquement un moyen de pilotage des données et par suite une expérience de monitoring des flux (d'énergie, d'eau, d'informations diverses), et dans ce secteur comme dans les autres mentionnés ici, l'illusion du déterminisme technologique doit être dénoncée et combattue

3 Parallèlement au développement de l'IA, la question de la souveraineté est aujourd'hui également questionnée à nouveaux frais par l'essor du numérique, voir à ce propos CERNA 2018.

[Greenfield 2013] ; pour autant quelle « politique des communs » engager pour rendre les différents modèles possibles compatibles avec une tradition de vie publique ? [Rochet 2014] Comment œuvrer à faire des villes intelligentes, « panoptiques électroniques » (*Electronic Panopticon*), mais également prodiges de soutenabilité urbaine, ces « écologies démocratiques » (*Democratic Ecologies*) capables d'encourager de nouvelles formes de civilité politique via l'accès généralisé aux données (*Open Data*) ? [Araya 2015 ; Auby & De Gregorio 2017]. À l'heure actuelle, ces questions demeurent ouvertes.

3. *Clarifier le type d'éthique requis pour ces besoins : le nécessaire travail de l'éthique sur elle-même*

Concernant le type d'éthique qui est requis pour traiter tous ces besoins actuels règne également une situation confuse que je voudrais clarifier.

Tout se passe en effet comme si les développements de l'IA étaient en train de faire travailler l'éthique sur elle-même. Le parallèle évoqué plus haut avec la bioéthique gagne à être approfondi : à partir de la fin des années 1960, les progrès considérables de la médecine (amélioration du diagnostic sous l'effet de nouvelles connaissances, réussite des premières greffes d'organes) puis l'essor de la biologie génétique ont engendré le développement de la réflexion bioéthique, qui s'est enrichie sous l'effet d'autres et nombreux développements technologiques tels que les perspectives ouvertes par la médecine d'« amélioration » (*enhancement*) [Allouche, Baertschi, Goffette *et al.* 2009] et par les thérapies génomiques, puis plus globalement encore sous l'effet des biotechnologies industrielles. Un pan entier de l'éthique philosophique, et jusqu'au concept nouveau de « bioéthique », ont émergé dans des circonstances d'un développement scientifique effectué dans le contexte d'émergences technologiques et de développement économique de marché, mettant en relation de manière originale et imprévue experts, usagers, firmes et institutions.

Il est plausible qu'à l'heure actuelle, sous l'effet des développements techniques de l'informatique et du numérique (développements qui concernent d'ailleurs aussi la médecine et la biologie), il en aille pareillement du côté de l'éthique appliquée à l'IA. Il est par exemple possible que l'éthique de l'IA (*Ethics of Artificial Intelligence*), c'est-à-dire un des secteurs de l'éthique informatique (*Computer Ethics*), permette l'apparition de développements technologiques offrant aux machines la possibilité d'assister les humains dans leurs choix de valeurs, jusqu'à pouvoir automatiser en grande partie le raisonnement éthique [Bostrom & Yudkowsky 2014]. L'éthique de l'IA pourrait dans ces conditions être assimilée à une branche de la science informatique, comme la bioéthique est parfois présentée comme une ramification de la médecine.

Que l'éthique travaille sur elle-même, cela signifie qu'elle est conduite à distinguer et prioriser les formes d'éthique adéquates aux tâches à accomplir. Or, à l'heure actuelle, en dépit des apparences, personne ne peut dire avec certitude quelle forme d'éthique est exactement adéquate au traitement approfondi des situations engendrées par le développement de l'IA dans la société. Clore trop rapidement ce débat pourrait même conduire à un appauvrissement préjudiciable à la cause que l'on veut promouvoir grâce à lui, à savoir, redonner voix, dans un monde en cours d'automatisation exponentielle, à un humain à la fois actif et responsable.

Un tel risque se dessine aujourd'hui du fait du primat croissant de l'éthique conséquentialiste qui, dans le contexte de l'économie de l'innovation, tend à dominer les situations d'évaluation des développements technologiques. Le conséquentialisme désigne une démarche qui détermine le caractère bon ou mauvais d'une intention ou d'une action en regard de leurs conséquences, c'est-à-dire, d'abord, de leurs effets observables et évaluables d'un point de vue donné, ce dernier pouvant varier. On se souvient que ce type de raisonnement éthique a été baptisé, dans une contribution à la visée essentiellement critique,

par G. E. M. Anscombe : dans son article « Modern Moral Philosophy » de 1958, Anscombe remarque que, faute de disposer d'un principe de jugement naturel ou d'origine théologique qui rendrait possible l'expérience personnelle du devoir, et parce qu'elle ne peut plus valider qu'une intention peut être *en elle-même* qualifiée par les termes « bien » ou « mal », la philosophie morale en est réduite à spéculer (et selon elle de manière incertaine) sur *les effets* de ces intentions ou actions. Si elle n'a pas été directement inventée et promue par l'utilitarisme, elle a tout de même préparée par lui, au fil d'une évolution qui part des recherches de John Stuart Mill et aboutit à celles de George Edward Moore *via* l'œuvre d'Henry Sidgwick ; est progressivement « acceptée pour être plutôt évidente l'idée que « l'action juste » signifie celle qui produit les meilleures conséquences possibles » [Anscombe 1958/2008 : 20].

L'approche conséquentialiste semble certes présenter d'incontestables avantages pour les cas de délibération provoqués par l'émergence des innovations : elle permet de construire par le raisonnement une réponse éthiquement satisfaisante, de manière toujours adéquate à une situation donnée et, chose très importante en matière de diffusion de l'innovation dans la société, sans jamais préjuger du résultat avant le raisonnement de l'usager. Or cela ne va pas sans poser un certain nombre de problèmes qui pourraient rendre cette approche discutable si on se limitait à elle.

D'une part, le conséquentialisme repose sur le postulat de la capacité qu'ont les agents moraux de concevoir avec clarté et distinction les conséquences de leurs intentions et de leurs actions. Or, dans le cas de toutes les innovations comme dans celui des développements de l'IA, ce postulat n'est pas avéré, en tout cas pas d'une manière absolue : en effet, nombre d'innovations n'ont pas été réalisées de la manière dont elles avaient été initialement conçues, et, une fois mises en marché, elles sont rarement utilisées conformément aux préconisations de leurs concepteurs, mais littéralement détournées par les usagers – l'IA, dans ses déclinaisons mises en marché, ne devrait pas échapper à cette logique. Sans que cela la désamorce complètement ou par principe, l'approche conséquentialiste se voit en quelque sorte défiée par les cas de sérendipité caractéristiques des situations d'innovation.

D'autre part, une des conséquences de cette adaptation désormais commune est que certains standards apparaissent comme des cadres presque incontournables de la réflexion, de manière sans doute illusoire. À cet égard, la manière dont on a traité une des innovations pour lesquelles, aujourd'hui, on invoque fréquemment la nécessité d'un traitement éthique de l'IA, à savoir, le véhicule autonome (VA), est éloquente. Le recours au conséquentialisme semble imposer l'argument dit du « dilemme du tramway », cette expérience de pensée particulière qui invite à des choix basés sur un calcul du coût létal, comme une sorte de référence obligée. Le succès du site du M.I.T. intitulé « Moral Machine »[4], plate-forme expérimentale en ligne conçue pour explorer les dilemmes moraux des véhicules autonomes, qui invite à « jouer » au dilemme du tramway appliqué à l'IA, repose en grande partie sur les séductions engendrées par ce type de calcul. Tout se passe donc comme si le VA, ce symbole de l'innovation de rupture qui guette aujourd'hui le monde de la mobilité sous l'effet de l'adoption généralisée de l'IA, devait être éthiquement pensé par référence au célèbre dilemme du tramway. Pourtant, si en apparence il permet de clarifier les options et de ce fait simplifie le choix éthique, à certains égards cet argument envenime la situation : comme il ne permet pas de dépasser la situation où le choix humain provoque un certain nombre de décès, le conséquentialisme (en tout cas sous la forme extrémiste du dilemme du tramway) s'avère contre-performant car vecteur de désespoir. Lorsqu'il adopte cette forme, ne se présente-t-il

4 http://moralmachine.mit.edu/hl/fr. Cette plate-forme a recueilli 40 millions de décisions en dix langues différentes et issues de 233 nations ou régions différentes [Awad *et alii* 2018].

pas, à l'égard des efforts humains pour sortir du cercle de la décision létale, comme un sophisme tragique ?

Aussi semble-t-il pertinent de tenter de pluraliser les formes du raisonnement éthique, au lieu de s'enfermer dans le seul paradigme conséquentialiste. La littérature de philosophie morale offre la ressource de trois autres formes possibles de raisonnement éthique : au conséquentialisme peut en effet se trouver associé le **déontologisme** (ensemble des éthiques qui privilégient l'expérience morale intériorisée ou devoir), **l'arétaïsme** (terme sous lequel se regroupent les éthiques des vertus) et **l'axiologisme** (ensemble des éthiques privilégiant une valeur cardinale, on peut ici convoquer les éthiques d'origine religieuse).

Le déontologisme peut être illustré par la figure de Kant qui, dans les *Fondements de la métaphysique des mœurs* (1785) et dans la *Critique de la raison pratique* (1788) s'est attaché à formuler la législation éthique qui conditionne la capacité humaine à forger un jugement moral autonome, c'est-à-dire un jugement basé sur la faculté d'orienter sa propre action par une réflexion désintéressée. Pour illustrer l'arétaïsme, on peut évoquer Platon, Aristote ou Augustin d'Hippone, tous auteur pour lesquels peut être dite éthiquement valable la décision ou l'action qui à la fois *repose sur* et *permet de développer* les vertus, à savoir, des éléments de caractère socialement féconds tels par exemple que le courage, la justice et la tempérance pour Platon, la magnanimité, la libéralité, selon Aristote, ou encore la bonté pour le christianisme (tout en disqualifiant les vices, à savoir, les éléments de caractère socialement néfastes tels que la lâcheté, l'avidité, l'impulsivité, l'intolérance, l'avarice, etc.). Enfin, l'axiologisme peut renvoyer aussi bien aux éthiques religieuses d'origine monothéistes (la volonté d'un Dieu unique étant la valeur suprême), et en ce cas on peut l'illustrer avec la figure de leurs prophètes, qu'aux éthiques philosophiques prônant par exemple la valeur supérieure de la nature (Hans Jonas apparaît comme un bon candidat pour l'éthique écologiste) ou celle de la nation (Ernest Renan).

En dépit du tropisme de l'éthique computationnelle dominant le paysage de l'IA, qui privilégie le conséquentialisme d'origine utilitariste, ce n'est pas une seule, mais bien quatre formes d'éthique qui existent. Toutes sont cohérentes, attestées dans la littérature, et apparaissent pertinentes pour la démarche visant à enrichir la contribution humaine aux choix éthiques face aux développements de l'IA. Dans notre situation contemporaine, pour toute tentative d'évaluation éthique, il serait donc sans aucun doute aussi utile que pertinent de formuler des arguments empruntés non pas à une seule, mais bien aux quatre familles de l'éthique. Pourtant, une telle tentative n'est guère aisée, elle est même très difficile, et c'est ce point que je voudrais souligner en conclusion.

Conclusion

Se lancer dans la recherche d'une éthique de ou pour l'IA s'impose aujourd'hui dans le contexte d'un contraste émotionnel puissant ressenti par nos contemporains, un contraste entre d'un côté l'enthousiasme général qui semble accompagner le développement des technologies (*machine/deep learning*, robotique…), et de l'autre les incertitudes ou inquiétudes éprouvées vis-à-vis des transformations de pratiques induit par ce développement. Mais de ce fait nous sommes également confrontés (et de manière potentiellement désagréable) à nos propres ambiguïtés vis-à-vis de la technologie. Qu'attendons-nous de l'IA ? Sans doute beaucoup trop, car au-delà des nouveaux et fascinants outils dont elle nous dote, nous voyons l'IA comme capable d'opérer une amélioration radicale de notre condition humaine. Évidemment, une telle illusion doit être combattue, mais ce qui rend le combat difficile est le fait que, dès le début, le développement de l'IA a été inclus dans un projet de société cohérent car en phase avec la perspective moderniste de la rationalité et de la science (esquissée par Descartes, validée par Newton et Leibniz). Ce projet de société, le

mathématicien Norbert Wiener (1894-1964), l'avait préparé dans les années 1940 en imaginant la cybernétique comme science du fonctionnement de l'esprit humain, et il a été esquissé en 1956 à la conférence de Dartmouth par John McCarthy, C.E. Shannon, Marvin L. Minsky et N. Rochester. Les organisateurs de la conférence de Dartmouth prévoyaient que soient abordées diverses questions autour de l'idée de machine pensante : comment simuler la pensée et le langage grâce à des règles formelles ? Comment faire penser un réseau de neurones ? Comment doter une machine de capacité d'apprentissage automatique ? Si au final la conférence de Darmouth n'a pas accouché de résultats de recherche immédiats, elle a lancé une impulsion décisive dont l'effervescence contemporaine sur l'IA est l'héritière, et permis de délimiter les contours d'une communauté de recherche autour des problématiques évoquées. Le style intellectuel des recherches autour de l'IA était né. C'est, si je puis écrire, ce style même qui met en question la possibilité de nourrir l'éthique de l'IA aux meilleures sources de la théorie morale.

En effet, qu'est-ce que l'éthique, entendue de manière générale ? La démarche éthique consiste à délimiter la responsabilité humaine et à permettre aux humains de savoir comment orienter leur liberté. Aussi, revendiquer l'éthique revient, pour un sujet, à assumer l'ambition d'être normatif. Est normatif ce qui émet des jugements de valeur, institue des règles et des principes d'action. L'ambition éthique est spécifique, car elle s'inscrit dans la perspective de la recherche d'une normativité qui n'est ni celle de la loi, ni celle de la coutume, ni celle de la déontologie d'une profession ou d'une corporation : si les diverses normativités, qu'elles soient juridique, sociologique ou éthique, peuvent parfois concorder, elles peuvent également diverger voire s'opposer. Le non-alignement des normativités contribue au fait que la démarche éthique (la recherche par un sujet d'un sens et de règles qui guident son action) assume une fonction de structuration de l'expérience intérieure, et, pour employer une métaphore, permet de rendre « consistante » la subjectivité, notamment via les épreuves morales, une épreuve pouvant être définie comme un problème vécu pour la résolution duquel il est nécessaire de mobiliser des ressources qu'on ne se sait pas posséder a priori.

L'éthique computationnelle, issue du projet cybernétique, concerne quant à elle l'échange des informations entre des émetteurs et peut-être à cet égard dite relationnelle plutôt que réflexive, et, du fait de son origine dans la discipline informatique, formelle plutôt que substantielle. Elle se distingue des autres formes d'éthique, car les formes déontologique, arêtaïque et axiologique caractérisent pour leur part, chacune à leur manière, l'approfondissement réflexif des conduites au nom de valeurs. Si, considérée en général, toute démarche éthique se noue à une expérience spécifique ou irréductible qui apparaît comme une épreuve qu'on peut dénouer au nom de valeurs, la démarche conséquentialiste qui préside à l'éthique computationelle apparaît comme une suite de raisonnements logiques sans épreuves ni valeurs. Une telle expérience fait-elle réellement sens pour un sujet humain ? Dans une certaine mesure sans doute, puisque l'éthique computationnelle, structurée par le raisonnement conséquentialiste permet d'espérer traiter efficacement les cas d'usage que nous soumet la nouvelle société algorithmique. Toutefois, dans une autre mesure, elle ne peut constituer une véritable éthique que si elle permet aux subjectivités de faire l'expérience de l'épreuve et du choix des valeurs, et de vivre celle du conflit des normativités et de la recherche de leur liberté, tant intérieure que publique.

Bibliographie :

- Allouche, Sylvie, Baertschi, Bernard, Goffette, Jérôme, et al., 2009 : *Enhancement. Ethique et philosophie de la médecine d'amélioration*, Paris, Éditions philosophiques J. Vrin.
- Anscombe, Gertrude Elisabeth Margaret, 1958/2008: "Modern Moral Philosophy", *Philosophy*, vol. 33, n^{o} 124. Trad. fr. *Klesis*, n°9-2008.

- Araya, Daniel (eds), 2015 : *Smart Cities as Democratic Ecologies*, New York, Palgrave Macmillan.
- Auby, Jean-Bernard, De Gregorio, Vincenzo (dir.), 2017 : *Données urbaines et* Smart Cities, Paris, Éditions Berger-Levrault.
- Awad, Edmond, *et alii*, 2018: "The Moral Machin Experiment", *Nature*, Volume 563, p. 59–64.
- Bostrom, Nick, Yudkowsky, Eliezer, 2014 : « The Ethics of Artificial Intelligence », in Ramsey, William, Frankish, Keith, eds., *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, 316-333.
- Brey, Philip, 2004 : « Ethical aspects of facial recognition systems in public places », *Journal of Information, Communication and Ethics in Society*, Vol. 2 Issue: 2, p. 97-109.
- Caseau, Yves, Soudoplatoff, Serge, 2016 : *La blockchain, ou la confiance distribuée*, Fondapol, Fondation pour l'innovation politique.
- Chamayou, Grégoire, 2013 : *Théorie du drone*, Paris, Éditions La Fabrique.
- CNIL 2017 : *Comment permettre à l'Homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, Commission Nationale Informatique & Libertés, décembre 2017.
- Commission de réflexion sur l'éthique de la recherche en sciences et technologies du numérique (CERNA), 2017, *Éthique de la recherche en apprentissage machine*, Rapport de recherche, consulté à l'URL http://cerna-ethics-allistene.org/digitalAssets/52/52472_CERNA_Ethique_de_la_recherche_en_apprentissage_machine.pdf le 20 mai 2019.
- Dumouchel, Paul, Damiano, Lisa, 2016 : *Vivre avec les robots. Essai sur l'empathie artificielle*, Paris, Éditions du Seuil.
- European Commission, 2019: *Ethics guidelines for trustworthy AI*, avril 2019, accessible à l'URL : https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, consulté le 10 septembre 2019.
- Germain, Éric, 2019 : « C'est quoi un « robot tueur » ? Dix années de débat sur les armes « autonomes » », *Revue Défense Nationale*, Mai 2019, n° 820, accessible à l'URL : http://www.defnat.com/e-RDN/vue-article.php?carticle=22056, consulté le 1er juin 2019.
- Google 2018 : « AI at Google: our principles », blog de Sundar Pichai, Directeur Général de Google : https://blog.google/technology/ai/ai-principles/, publié le 7 juin 2018, consulté le 27 mai 2019.
- Greenfield, Adam, 2013 : *Against the Smart City (The City is Here for You to Use, 1)*, Kindle Edition.
- Guchet, Xavier, 2016 : *La Médecine personnalisée. Un essai philosophique*, Paris, Editions Les Belles Lettres.
- Lin, Patrick, 2015 : « Why Ethics Matters For Autonomous Cars », in Maurer, Markus,*et al*. (eds), *Autonomes Fahren*, Berlin-Heidelberg, Springer, p. 69-85.
- Mallard, Alexandre, Méadel, Cécile & Musiani, Francesca, 2014 : *The Paradoxes of Distributed Trust: Peer-to-Peer Architecture and User Confidence in Bitcoin*. Journal of Peer Production, 2014, accessible à l'URL : http://peerproduction.net/issues/issue-4-value-and-currency/peer-reviewed-articles/theparadoxes-of-distributed-trust/, consulté le 1er juin 2019.
- Marciano, Avi, 2019 : « Reframing Biometric Surveillance: From a Means of Inspection to a Form of Control », *Ethics and Information Technology*, 21 (2), p. 127-136.
- Microsoft, 2018 : *Microsoft AI principles, Designing AI to be trustworthy requires creating solutions that reflect ethical principles that are deeply rooted in important and timeless values*, consultable à l'URL : https://www.microsoft.com/en-us/ai/our-approach-to-ai, consulté le 10 septembre 2019.

- Perrin, Jérôme, 2019 : « Peut-on élaborer une politique éthique du véhicule autonome », *Revue Défense Nationale*, Mai 2019, n° 820, Cahiers de la RDN, n°108.
- Picon, Antoine, 2013 : *Smart Cities : Théorie et critique d'un idéal auto-réalisateur*, Paris Éditions B2.
- Ricœur, Paul, 1995 : « Le concept de responsabilité. Essai d'analyse sémantique », dans *Le Juste, I*, Paris, Éditions Esprit, p. 41-70.
- Rochet, Claude, 2014 : *Les villes intelligentes, enjeux et stratégies pour de nouveaux marchés. Le programme MUST (Management of Urban Smart Territories)*, étude réalisée en septembre 2014 avec le concours de CERGAM, CESAMES, et du Service de coordination à l'intelligence économique (Ministère français des finances et des comptes publics), Paris, ESCP Europe, Business School.
- Schlosser, Markus, 2015 : « *Agency* », in *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, accessible à l'URL : https://plato.stanford.edu/entries/agency/, consulté le 22 mai 2019.
- Shelley, Mary, 1818 : *Frankenstein ou le Prométhée moderne*, trad. J. Ceurvorst, Paris, Marabout, 1978.
- Toronto, 2018: *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems* (Amnesty International, Access Now, Human Rights Watch, Wikimedia Foundation *et alii*), téléchargeable en PDF à l'URL : https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf, consulté le 10 septembre 2019.