Éthique et IA : analyse et discussion

Catherine Tessier

Éthique et IA : analyse et discussion

C. Tessier[1]

[1] ONERA/DTIS, Université de Toulouse

catherine.tessier@onera.fr

Résumé

La profusion de documents ainsi que d'instances créées pour traiter de « l'éthique de l'IA » amène à s'interroger sur les raisons pour lesquelles l'IA est devenue, depuis quelques années, un objet particulier d'attention, pourquoi cet objet est spécifiquement regardé sous un angle dit « éthique » et de quelle éthique il s'agit. L'examen des textes européens, de l'UNESCO, de l'OCDE et de la déclaration de Montréal révèle notamment une interprétation sémantique des notions qui peut prêter à confusion, et des postulats susceptibles d'affecter les réflexions. Des tensions et paradoxes peuvent être mis en évidence, que nous illustrons en particulier sur le principe du contrôle humain. Nous insistons en conclusion sur les risques de dévoiement de l'éthique et la nécessité d'une véritable réflexion éthique accompagnant les évolutions techniques et applicatives en matière d'IA.

Mots-clés

Intelligence artificielle, éthique, tensions, contrôle humain.

Abstract

The high number of documents as well as bodies created to deal with « the ethics of AI » leads us to wonder why AI has become, in recent years, a particular object of attention, why AI is specifically looked at from the « ethics » point of view and which ethics is at stake. A review of European, UNESCO, OECD documents and of the Montreal Declaration reveals a semantic interpretation of notions that can lead to confusion, and some postulates that can be misleading. Tensions and paradoxes can be highlighted, which we illustrate in particular on the principle of human control. In conclusion, we insist on the risks of misuse of ethics and the need for a true ethical reflection going with the technical and applicative evolutions in AI.

Keywords

Artificial Intelligence, Ethics, Tensions, Human control

1 Introduction

Le rapport annuel *Artificial Intelligence Index 2019* [14]-(page 273) de l'université de Stanford recensait cinquante-huit documents, toutes sources confondues (organisations officielles et gouvernementales, universités, sociétés savantes, industries, *think tanks*) associant « intelligence artificielle (IA) » et « éthique ». L'observatoire de l'intelligence artificielle créé par l'OCDE (Organisation de coopération et de développement économiques) propose sur son site une base de données interactive des documents de politiques et initiatives en matière d'IA [22], dont ceux qui traitent d'« éthique ». Quant au rapport de l'Université de Harvard [11], il analyse trente-six documents traitant de principes pour l'IA : les principes les plus souvent invoqués sont la protection de la vie privée, la répartition des responsabilités (*accountability*), la sûreté et la sécurité, la transparence et l'explicabilité, l'équité (*fairness*) et la non-discrimination, le contrôle humain, la responsabilité des professionnels, le respect des valeurs fondamentales.

La profusion de documents ainsi que d'instances créées pour traiter de « l'éthique de l'IA » amène à s'interroger sur les raisons pour lesquelles l'IA est devenue, depuis quelques années, un objet particulier d'attention, pourquoi cet objet est spécifiquement regardé sous un angle dit « éthique » et de quelle éthique il s'agit. L'examen des textes révèle notamment une interprétation sémantique des notions qui peut prêter à confusion, et des postulats susceptibles d'affecter les réflexions. Des tensions et paradoxes peuvent être mis en évidence, que nous illustrerons en particulier sur le principe du contrôle humain. Nous insisterons en conclusion sur les risques de dévoiement de l'éthique et la nécessité d'une véritable réflexion éthique accompagnant les évolutions techniques et applicatives en matière d'IA.

L'analyse qui suit est fondée principalement sur les textes européens émanant du Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle, du Parlement et de la Commission ; sur le texte provisoire de l'UNESCO rédigé par le Groupe d'experts *ad hoc* ; sur la Déclaration de Montréal ; sur les textes de l'OCDE.

2 De quoi parle-t-on ?

2.1 Intelligence artificielle

Le vocable « intelligence artificielle », mal choisi [17], fait l'objet, y compris chez les scientifiques, de dérives de langage qui amènent à personnifier l'IA et à attribuer aux logiciels des caractéristiques équivalentes à celles d'un être vivant : « *une IA* fait ceci ou cela ». On observe également l'emploi de « IA » en tant que synonyme de « logiciel », « machine » ou « système », même si ceux-ci comprennent

des techniques qui ne relèvent pas de l'IA.

L' « autonomie » d'un « agent » ou d'un robot crée également des confusions, des fantasmes et des erreurs de raisonnement, y compris au sein d'instances internationales de négociation (par exemple les négociations à la Convention sur certaines armes classiques (CCAC) à Genève au sujet des systèmes d'armes létaux dits « autonomes »). De même le terme « déléguer », employé pour exprimer le fait que certaines fonctions habituellement dévolues à un être humain sont programmées dans une machine [15] sous-entend que l'être humain transfère une partie de ses responsabilités à la machine, celle-ci étant ainsi susceptible d'être dotée d'une existence morale ou juridique.

En outre, si « intelligence artificielle » est définie correctement en préambule des documents étudiés, c'est-à-dire en expliquant la diversité des approches rassemblées sous ce vocable (comme figurée sur le « diamant de l'IA » de l'AFIA [4]), le vocable est largement entendu dans le corps des textes comme un synonyme de « apprentissage machine ». Cela apporte d'autant plus de confusion que les recherches se concentrent actuellement sur l'hybridation d'approches statistiques et symboliques, ces dernières relevant de l'intelligence artificielle dans son sens premier.

Ces ambiguïtés dans le vocabulaire et les définitions contribuent à créer plusieurs écueils :

- L'objet du discours n'est pas clair : dans les textes relatifs à l'« éthique de l'IA », est-il question de l'ensemble des approches relevant de l'IA ou spécifiquement de celles qui sont fondées sur l'apprentissage machine ?
- L'anthropomorphisation de l'IA peut conduire à une surestimation des possibilités et des risques [28] : les machines pourraient ainsi « décider par elles-même » ou « prendre des initiatives » comme par exemple, pour un véhicule dit « autonome », « choisir » de renverser telle personne plutôt que telle autre ;
- Le vocabulaire employé peut laisser croire que les machines et les programmes pourraient être mis sur le même plan moral que l'être humain, voire être des « machines morales » (en particulier lorsque des connaissances ou des comportements relevant de concepts de la morale ou de l'éthique normative y sont modélisés [9]). Cela peut être renforcé par des applications qui brouillent les repères en faisant passer des machines pour des humains (imitation de l'aspect physique, de la voix, d'interactions sociales).

En ce qui concerne les deux derniers points, le texte provisoire de l'UNESCO [7]-(alinéa 126) indique que : *« Les États membres devraient instaurer des politiques visant à sensibiliser à l'anthropomorphisation des technologies d'IA, notamment en ce qui concerne les termes utilisés pour les désigner, et évaluer les manifestations, les conséquences éthiques et les possibles limites de ce phénomène [...] »*. Parmi ces conséquences figurent d'autres ambiguïtés concernant les termes qui qualifient les systèmes d'IA dont il est question dans les documents relatifs à l'« éthique de l'IA ».

2.2 Éthique

L'examen des documents produits par les groupes d'experts, comités, instances nationales ou internationales au sujet de l'« éthique de l'IA » montre que la possibilité de s'interroger sur le fait de ne pas développer ou de ne pas utiliser des systèmes visant à automatiser les processus de décision ou fondés sur l'apprentissage machine n'est évoquée que de manière très marginale. De tels systèmes existent ou vont exister et il s'agit, d'une certaine manière, de les cautionner, en rappelant que des principes doivent être respectés, en énonçant des précautions à prendre, et en suggérant une approche fondée sur une évaluation des risques. Cette démarche que Marc Hunyadi qualifie de *« petite éthique »* s'inscrit dans une logique du fait accompli, où chacun dispose d'une liberté de plus en plus limitée de choisir de ne pas posséder ou de ne pas utiliser certains objets, et qui construit petit à petit *« des modes de vie imposés par personne en particulier et auxquels tout le monde adhère »* [13]. De plus, les questionnaires d'auto-évaluation qui sont proposés de manière institutionnelle [24] ou par des organisations privées, ou les comités *ad hoc* qui sont constitués, risquent de généraliser un blanchiment éthique (*ethics washing*) en promouvant une « conformité éthique » dont la valeur et le sens peuvent être discutables. Il est question en effet d'*« IA éthique »* [8], de *« conformité éthique des systèmes d'IA »* [7], de *« certificat européen de conformité éthique »* [27], voire d'*« éthicité »* des systèmes [16]. Ceci appelle les remarques suivantes :

- Un objet, un programme ou une technique ne peut pas être « éthique » en soi et ne peut être qualifié d'« éthique ». L'adjectif « éthique » (par définition [1] : qui concerne la morale) ne peut être associé qu'à une démarche, une délibération, une réflexion, une question, un principe, une valeur, etc.
- De même une conformité ne peut être « éthique » et il ne suffit pas de dire ce qu'il convient de faire ou ne pas faire. Ce qui relève de l'éthique est instable, singulier, et a à voir avec des dilemmes qui justement vont conduire à des solutions partiellement non conformes, qui ne vérifient pas toutes les propriétés (voir 4). La conformité dont il s'agit est une conformité technique à certaines exigences, énoncées dans un cahier des charges et vérifiées, y compris les éventuels compromis, par des simulations, des expérimentations, des campagnes de vérification, des processus d'homologation.
- Le concept d'« éthique par conception » (*"ethics by design"*) [27], calqué sur celui de respect de la vie

1. TLF et Larousse

privée dès la conception (*"privacy by design"*)[2], et compris comme l'intégration de principes allant au-delà des exigences légales dans la conception de systèmes d'IA [11], se heurte aux deux premières remarques. En particulier, *« éthique et état de droit dès la conception »* signifient dans [8]-(alinéas 98 à 101) : conformité aux normes, explicabilité, essais et validation, ce qui ne relève pas *a priori* de la réflexion éthique. Il existe de plus une ambiguïté sur l'expression française « éthique par conception » où « éthique » peut être compris en tant qu'adjectif – l'objet serait « éthique » (*ethical*) par conception – ou en tant que substantif (*ethics*) – de l'éthique serait prise en compte dès la conception de l'objet. Les auteurs de [28] indiquent qu'il serait préférable de concevoir des machines qui nous aident à agir mieux d'un point de vue éthique plutôt que d'envisager des machines comme des agents moraux ou se comportant conformément à des règles morales.

D'autre part, l'expression *« IA digne de confiance »* ou *« IA de confiance »* (*"trustworthy AI"*) qui selon l'Europe est définie par les trois caractéristiques : *« IA licite »*, *« IA éthique »* et *« IA robuste »* [8] est problématique. La confiance ne se décrète pas et une machine ou un système ne peut pas porter, en soi, la confiance. C'est bien l'expérience d'une personne qui utilise un système, l'examen de la manière dont il a été conçu et les garanties démontrées de conformité technique qui sont fournies qui vont amener cette personne à avoir confiance, ou non, dans ce système pour répondre à ses besoins. Comme l'affirme J. Bryson [5], *"No one should trust IA"*.

3 Les postulats

Les textes étudiés se fondent explicitement ou implicitement sur des postulats qui peuvent être discutables et occulter des éléments de réflexion. Nous relevons trois de ces postulats.

3.1 Les systèmes d'IA sont inéluctables

Aucun des documents étudiés n'envisage que les systèmes d'IA fassent l'objet de questionnements relatifs à leur existence même, à leurs raisons d'être. C'est une approche conséquentialiste fondée sur les risques et les précautions qui est adoptée, accompagnée de la nécessité d'un *« contrôle humain »* (voir 5) des systèmes d'IA, en particulier pour ceux qui sont estimés *« à haut risque »* [10, 27]. Le document de l'UNESCO [7] envisage cependant des interrogations sur l'utilisation des systèmes d'IA, en notant que celle-ci revêt un *« caractère facultatif »* (alinéa 20) et qu'une analyse devrait être menée pour évaluer si *« l'adoption de l'IA est appropriée »* (alinéa 58).

2. En Europe, il s'agit de concevoir les systèmes qui traitent des données à caractère personnel de manière conforme au Règlement général pour la protection des données (RGPD).

3.2 Les systèmes d'IA contribuent au bien-être

Les systèmes d'IA sont conçus dans l'*« objectif d'améliorer le bien-être et la liberté des êtres humains »*, constituent *« un moyen prometteur d'accroître la prospérité humaine, en renforçant ainsi le bien-être individuel et de la société ainsi que le bien commun »*, sont *« susceptibles d'apporter des avantages considérables aux individus et à la société »* [8]. L'IA « promet d'améliorer le bien-être des individus » [21]. *« Le développement et l'utilisation des systèmes d'intelligence artificielle doivent permettre d'accroître le bien-être de tous les êtres sensibles »* [1]-(principe 1).
L'Organisation mondiale de la santé mentionne toutefois que la notion de bien-être est multidimensionnelle, comprend des éléments subjectifs, culturels, et n'a pas de définition claire [25]. Selon l'Institut national de la statistique et des études économiques (Insee), contribuent au bien-être : les conditions de vie (logement, contraintes financières), la santé physique et émotionnelle, les liens sociaux, la sécurité, les risques psychosociaux au travail, les revenus, la composition du logement, l'âge, le diplôme [3].

3.3 Les systèmes d'IA sont une solution à tout

Dans le prolongement du postulat précédent, les systèmes d'IA peuvent contribuer à *« promouvoir l'égalité entre les sexes et lutter contre le changement climatique, rationaliser notre utilisation des ressources naturelles, améliorer notre santé, notre mobilité et nos processus de production, et nous aider à surveiller nos progrès par rapport à des indicateurs de durabilité et de cohésion sociale »* [8] ; il peuvent favoriser *« le renforcement des capacités humaines et le renforcement de la créativité humaine, l'inclusion des populations sous-représentées, la réduction des inégalités économiques, sociales, entre les sexes et autres »* [21] ; ils peuvent *« amélior[er]les conditions de vie, la santé et la justice, en créant de la richesse, en renforçant la sécurité publique ou en maîtrisant l'impact des activités humaines sur l'environnement et le climat »* [1].
A contrario, les contributeurs à l'atelier Quality of AI [17] de l'ERCIM (European Research Consortium for Informatics and Mathematics) soulignent que l'IA est souvent largement surestimée, mais qu'il est difficile de décrire tout ce qu'elle permet de réaliser sans laisser à penser qu'elle constitue une solution universelle. À titre d'exemple, les motivations qui justifient le déploiement de véhicules à conduite automatisée – amélioration de la sécurité routière, fluidification du trafic, réduction de la dépense énergétique, accès à la mobilité en particulier en zones rurales – sont en réalité *« peu documentées »* et assorties de fortes incertitudes [2]-(pages 20–22).

4 Tensions et paradoxes

4.1 Tensions entre principes

Les principes et exigences énoncés dans les documents ne peuvent pas être simultanément satisfaits, des compromis sont donc nécessaires [20]. Ces compromis, bien qu'ils

constituent justement l'objet de la réflexion éthique, sont évoqués de manière très succincte, par exemple : *« Des tensions pourraient survenir entre les principes [...], pour lesquelles il n'existe pas de solution unique. [...] Il faut [...] aborder les dilemmes et arbitrages éthiques selon une réflexion raisonnée et fondée sur des éléments probants. [...] Il pourrait toutefois exister des situations dans lesquelles aucun arbitrage acceptable du point de vue éthique ne peut être déterminé »* [8]-(alinéa 54) ; *« Si toutes les valeurs et tous les principes [...] sont souhaitables en soi, dans tout contexte réel, il y a inévitablement des compromis à faire, ce qui exige de procéder à des choix complexes concernant la hiérarchisation des contextes, sans pour autant compromettre d'autres principes ou valeurs »* [7]-(alinéa 11).

Quelques exemples de tensions sont proposés ci-dessous :

- Transparence / sécurité
  La transparence, l'explicabilité et la prédictibilité des systèmes d'IA peuvent présenter l'inconvénient d'une moindre sécurité et de possibles dérives d'usages si ces propriétés sont promues par l'ouverture des algorithmes voire des codes [3]. D'autre part, la transparence doit être évaluée au regard de la préservation de la propriété industrielle.
- Précision / protection de la vie privée
  Un système d'IA fondé sur l'analyse de données est d'autant plus précis et pertinent (*accurate*) que ces données sont précises, variées, riches et peuvent discriminer des situations particulières, voire rares, ce qui peut entrer en conflit avec la protection de la vie privée et des données à caractère personnel (données de santé ou de surveillance par exemple), voire la préservation des droits fondamentaux dans le cas des systèmes de reconnaissance faciale [18].
- Précision / préservation de l'environnement
  La précision d'un système d'IA fondé sur l'analyse de données nécessite de grands ensembles de données dont la collecte, le stockage et l'exploitation sont susceptibles d'avoir un fort impact sur l'environnement.
- Performance / Autonomie humaine
  Un système d'aide à la décision ou un système « autonome », conçu pour aider l'être humain ou le remplacer dans certaines tâches, est susceptible de porter atteinte à l'autonomie humaine, en influençant la décision de la personne, voire en s'y substituant. D'autre part, l'augmentation des capacités de ces systèmes, de leur pertinence et de leur fiabilité peut conduire à une dégradation, voire à la perte, de certaines compétences ou expertises humaines.

En outre et de manière paradoxale, les systèmes d'IA doivent être conçus de manière à respecter des principes et dans le même temps constituent une menace pour ces mêmes principes, ou bien être conçus pour un objectif qu'ils contribuent également à mettre en danger. Ainsi ils doivent être conçus dans le respect des droits fondamentaux et sont susceptibles de menacer ces droits ; ils peuvent améliorer le bien-être et abaisser la qualité de vie ; réduire les inégalités et les exacerber ; renforcer les capacités humaines et contraindre les choix des individus et des groupes ; renforcer la sécurité et ouvrir de nouvelles brèches de sécurité ; contribuer à lutter contre le changement climatique et affecter les écosystèmes, l'environnement et le climat [1, 8].

4.2 Équité et biais

Le principe d'équité (*fairness*) peut faire référence aux notions d'impartialité, d'égalité, de non-discrimination et de justice et suppose un idéal d'égal traitement des individus ou des groupes [24].

« Les biais et la discrimination sont des risques inhérents à toute activité sociétale ou économique » [10], cependant il est demandé aux acteurs de l'IA de *« réduire au maximum et éviter de renforcer ou de perpétuer des biais sociotechniques inappropriés basés sur les préjugés liés à l'identité »* [7]-(alinéa 29), de corriger les biais éventuels [10], de veiller à l'absence de *« biais injustes »* [8].

Dans le même temps il est demandé de réfléchir à la définition de l'équité [24]. Il semble en particulier nécessaire de situer la définition d'une propriété d'un logiciel ou des résultats qu'il est susceptible de fournir par rapport à la notion d'équité dans le sens commun. On peut constater d'abord que la nature n'est pas équitable en soi – *"unfairness is natural"* [4], que la société véhicule de nombreux biais et que les êtres humains, consciemment ou non, ont des comportements discriminatoires. Que signifie alors de réduire les biais ou d'éviter de les renforcer dans les systèmes d'IA, sous-entendu essentiellement ceux qui sont fondés sur l'analyse de données ?

On pourrait se demander ce que serait un objet logiciel sans biais, voire « neutre » et si des résultats de calcul qui seraient moralement neutres, équitables, seraient adaptés à la société ou à la nature. D'autre part, comment formaliser sous forme mathématique, donc programmable, un raisonnement ou une décision « juste » ou « équitable » ? Il semble que ces questions ne puissent être envisagées dans l'absolu : il convient de s'interroger sur la raison d'être et les objectifs de l'utilisateur du système d'IA ainsi que les valeurs qu'il souhaite promouvoir, et comment ces objectifs et valeurs orientent la conception du système. Par exemple, un processus automatisé de sélection de CV pour une embauche pourrait être fondé sur un tirage au sort ou sur l'historique des profils des personnes qui ont « réussi » au poste concerné. La première méthode, qui ne nécessite pas de système d'IA, peut être considérée – si toute personne a la possibilité de présenter son CV – comme

3. Softbank Robotics Webinar on Responsible Robotics and AI : Concrete solutions, Feb. 2021

4. J. Bryson, Softbank Robotics Webinar on Responsible Robotics and AI : Concrete solutions, Feb. 2021

« sans biais », mais a de grandes chances d'être inadaptée. La seconde est susceptible de perpétuer l'embauche de personnes ayant toujours les mêmes caractéristiques, sauf à diversifier la notion de « réussite », qui dépend des valeurs que l'organisation qui cherche à recruter veut renforcer grâce à cette embauche.

Remarque : Fairlearn [5] n'utilise pas le terme de « biais » et définit l'équité sur la base de deux types d'impacts des systèmes d'IA sur les personnes : préjudices d'affectation et préjudices de qualité de service.

5 Exemple : le contrôle humain

Il y a un consensus international sur le principe de « contrôle humain » des systèmes d'IA, qui se traduit dans les textes par des *« garanties et des mécanismes, tels que l'attribution de la capacité de décision finale à l'homme, qui soient adaptés au contexte et à l'état de l'art »* [21], le fait de pouvoir décider de ne pas utiliser un système d'IA afin de conserver des niveaux de jugements humains, ou d'assurer la possibilité que la décision de l'humain prime sur celle calculée par le système [24]. Pour les applications dites à « haut risque » (comportant des risques d'atteinte aux individus ou à la société), l'Europe préconise une garantie de *« participation adéquate de l'être humain »* [10], de supervision humaine à tout moment, et une reprise en main humaine quand nécessaire [27], et *« qu'à tout moment, une personne humaine ait la possibilité de corriger [le système], de l'interrompre ou de [le] désactiver en cas de comportement imprévu, d'intervention accidentelle, de cyberattaques, d'ingérence de tiers dans une technologie fondée sur l'IA ou d'acquisition par des tiers d'une telle technologie »* [26].

Cette notion de « contrôle humain » reste cependant floue en particulier parce qu'elle englobe plusieurs types d'interventions humaines : en effet, elle peut concerner le fait qu'une personne ou une organisation humaine décide d'utiliser ou non le système d'IA, la nature des décisions qui restent dévolues à l'humain, la supervision, les possibilités de reprise en main, les validations humaines des résultats fournis.

Par ailleurs se pose la question de l'évaluation de la présence du contrôle humain : comment et par qui cette évaluation est-elle réalisée ? Que signifie techniquement la garantie de supervision humaine « à tout moment » ? Comment garantit-on que l'intervention humaine est pertinente ?

5.1 Un paradoxe

Les raisons pour lesquelles on souhaite automatiser des fonctions décisionnelles dans le cadre d'une application ou d'ensembles d'applications sont multiples, par exemple : les tâches à réaliser dépassent les capacités humaines (le contexte demande par exemple d'envisager une combinatoire élevée ou un espace de recherche de solutions très grand) ; elles mettent en cause la sécurité ou la santé de l'humain (le contexte est dangereux ou hostile) ; l'automatisation est plus économique ; l'automatisation est plus sûre (elle permet de pallier l'erreur humaine).

Il y a donc un paradoxe entre les raisons qui motivent l'automatisation et le fait d'exiger un contrôle humain des fonctions automatisées : se pose en effet la question de la capacité de l'humain à exercer effectivement ce contrôle. De plus, la notion de contrôle humain sous-entend que le point de vue de l'humain est pertinent et correct, et qu'il doit prévaloir sur les résultats des calculs de la machine. Enfin, le contrôle humain nécessite qu'il y ait effectivement un humain présent et disponible – par exemple il est indiqué dans [23]-(section 9) que les services publics européens doivent être largement fondés sur des systèmes numériques à base d'IA, mais que le recours à un interlocuteur humain doit toujours être possible.

5.2 Limites du contrôle humain

L'humain doit disposer d'informations et de temps, qui soient compatibles avec le contrôle à exercer. En particulier, l'humain ne peut pas être considéré comme le recours ultime dans n'importe quelle situation ou quand les fonctions automatisées « ne savent pas faire ». Par exemple, il est illusoire d'envisager le transfert du contrôle de la conduite d'un véhicule « autonome » des automatismes vers l'utilisateur si celui-ci, comme on le voit dans certaines publicités de constructeurs automobiles, est occupé à d'autres activités : une bonne conscience de situation, incluant prédiction et anticipation, est indispensable pour élaborer des décisions et des actions adaptées. Même la procédure d'arrêt en sécurité (*"stop button"*) [24] que l'être humain pourrait engager est complexe à envisager de manière opérationnelle en toutes circonstances.

En outre, les automatismes altèrent les mécanismes de contrôle classiquement utilisés par l'humain : moindre engagement dans la tâche, augmentation de la divagation attentionnelle, moindre aptitude à détecter des erreurs. En particulier s'il est novice, fatigué ou stressé, l'humain est susceptible de se reposer sur ce que préconise la machine et d'être ainsi enfermé dans des choix restreints ou erronés. Enfin, le manque d'informations ou au contraire un flot trop abondant d'informations, ou les schémas que l'humain a en tête, peuvent entraîner une mauvaise compréhension du comportement de la machine ou des résultats qu'elle propose, et entraîner des décisions humaines inadaptées.

5.3 Envisager un partage de l'autorité

Le fait d'assurer la possibilité que la décision de l'humain prime sur celle calculée par le système d'IA, qui est l'une des options de la surveillance humaine [24] suppose que l'humain est infaillible.

Il ne s'agit pas d'opposer l'humain et la machine ou les logiciels, mais de répartir les bonnes compétences aux bons

5. A Python package to assess and improve fairness of machine learning models : `https://github.com/fairlearn/fairlearn`

endroits dans le cadre d'une approche système incluant les mécanismes humains mis en jeu, et en *analysant les besoins* : la machine doit être conçue pour aider ses utilisateurs et remplir un service bien identifié en préservant l'essence même de ce qui est important pour prendre des décisions et en endosser la responsabilité. D'un point de vue technique, des critères concrets permettant de spécifier et de vérifier la façon dont la machine permet à l'humain d'exercer ses mécanismes de contrôle doivent être définis. En outre il faut s'interroger sur ce que seront les capacités cognitives de demain, certaines capacités étant susceptibles de décroître pendant que d'autres se développent ; il faudra certainement adapter les systèmes à ces capacités différentes.

5.4 La question de l'annotation

Un type particulier de « contrôle humain » est la nécessaire annotation ou transcription des données pour alimenter les systèmes d'apprentissage. Si l'Europe recommande la mise en place d'outils souverains en la matière, dans le respect des législations [23], la question de l'annotation des données et de la transcription d'échanges verbaux n'apparaît pas en tant que préoccupation éthique dans les documents étudiés. Pourtant, ce contrôle humain est largement effectué par des « micro-travailleurs » précaires, sous-rémunérés et dépourvus de couverture sociale, ou bien par des employés de sous-traitants des grandes entreprises du numérique exposés en permanence à des données personnelles sensibles ou des à propos dérangeants. Il faut certainement s'interroger sur la tension entre performances des logiciels fondés sur les données et dignité et intégrité des personnes qui contribuent à ces performances.

6 Conclusion

6.1 Des risques de dévoiement de l'éthique

On ne peut pas laisser penser qu'« une IA éthique » serait possible, et que cela consisterait à vérifier une conformité à des critères et à des normes, traduits en exigences techniques. En effet les instruments normatifs [7], les standards (IEEE P7000™ et suivants), les grilles d'évaluation (auto-évaluation « éthique » pour les projets européens), les audits éthiques [20], qui relèvent essentiellement d'une bonne gestion et de bonnes pratiques, présentent le risque d'être substitués à une véritable réflexion éthique, permanente et toujours en chantier. Par exemple, s'il est nécessaire de vérifier la conformité d'un projet d'identification biométrique [6] au Règlement général pour la protection des données (RGPD), il est tout aussi nécessaire de réfléchir aux raisons pour lesquelles ce projet est mené et, et fur et à mesure des performances constatées, de se demander s'il faut le poursuivre et dans quelles directions, et quels dilemmes se présentent.

Le risque est un dévoiement de l'éthique qui, *via* des labels ou des certificats, pourrait à la fois donner bonne conscience et servir d'argument de vente. En ce sens, *une éthique de l'éthique de l'IA* [28] est à construire.

6. https://www.cnil.fr/fr/biometrie

6.2 Une indispensable réflexion éthique

« Le premier danger que présente le développement de l'intelligence artificielle consiste à donner l'illusion que l'on maîtrise l'avenir par le calcul. [...] Mais dans les affaires humaines, demain ressemble rarement à aujourd'hui, et les nombres ne disent pas ce qui a une valeur morale, ni ce qui est socialement désirable. » [1]

Outre une approche conséquentialiste fondée sur l'analyse des risques, il s'agit également de questionner l'existence même de l'objet (logiciel, robot), sa raison d'être, et se demander si cet objet est désirable et au nom de quelles valeurs. La référence étant des moyens existants ou l'être humain, l'objectif peut être de faire effectuer des tâches plus rapidement, à plus grande échelle, de manière plus précise, plus sûre, voire plus « inventive » ; de réduire des coûts ; de proposer des solutions plus simples, plus commodes, plus ludiques. On peut aussi interroger la pertinence de ces critères : pourquoi vouloir aller plus vite, etc., et éloigner toujours plus l'être l'humain, et ce de manière paradoxale car le contrôle humain est considéré comme impératif ?

Une analyse des besoins, à mettre en regard de l'évolution des capacités techniques, permettrait de distinguer usages et technologies, d'envisager les usages problématiques et les glissements insidieux vers de tels usages, d'interroger la légitimité d'utiliser certaines techniques, d'identifier des nouveaux besoins susceptibles d'être créés de toutes pièces, ou de questionner l'utilisation des technologies pour traiter des effets de dérives plutôt que de remédier aux dérives elles-mêmes (par exemple la relecture d'articles par les pairs assistée par « IA » pour faire face à l'inflation de propositions de publications [6]). Il s'agirait plus d'« embarquer » l'éthique dans la recherche et l'ingénierie relatives à l'IA plutôt que dans l'IA elle-même, c'est-à-dire la considérer pour ce qu'elle est, un processus de réflexion continu qui concerne des tensions entre principes et des questions sans « bonne solution » [19]. Il pourrait être imposé par exemple une discussion éthique dans les articles scientifiques [12]. Ce n'est pas l'« IA » qui va évoluer toute seule vers de nouvelles capacités et applications, comme peut le laisser croire la façon dont on personnifie cet ensemble de techniques, mais bien les humains qui doivent décider collectivement de ce qu'il convient ou non de faire, en analysant si ces progrès technologiques vont nécessairement vers un progrès moral.

Liens d'intérêt

L'autrice a été membre du Groupe d'experts *ad hoc* de l'UNESCO pour l'élaboration d'un avant-projet de recommandation sur l'éthique de l'intelligence artificielle [7].

Références

[1] La Déclaration de Montréal pour un développement responsable de l'Intelligence Artificielle, 2018. https://www.declarationmontreal-iaresponsable.com/la-declaration.

[2] Développement des véhicules autonomes - Orientations stratégiques pour l'action publique. Ministère de la transition écologique, 2018. https://www.ecologie.gouv.fr/developpement-des-vehicules-autonomes-orientations-strategiques-laction-publique.

[3] M.-H. Amiel, P. Godefroy, and S. Lollivier. Qualité de vie et bien-être vont souvent de pair. Technical report, Insee, 2013. https://www.insee.fr/fr/statistiques/1281414.

[4] Association française pour l'Intelligence Artificielle. Domaines de l'IA. https://afia.asso.fr/domaines-de-lia/.

[5] J. Bryson. AI & Global Governance : No One Should Trust AI. United Nations University, Centre for Policy Research, 2018. https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html.

[6] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi. AI-assisted peer review. *Humanities and Social Sciences Communications*, 8 :25, 2021. https://doi.org/10.1057/s41599-020-00703-8.

[7] Groupe d'Experts ad hoc (GEAH) de l'UNESCO. Avant-projet de recommandation sur l'éthique de l'Intelligence Artificielle, 2020. https://unesdoc.unesco.org/ark:/48223/pf0000373434_fre.

[8] Groupe d'Experts Indépendants de Haut Niveau sur l'Intelligence Artificielle. Lignes directrices en matière d'éthique pour une IA digne de confiance. Commission européenne, 2019. https://op.europa.eu/fr/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/prodSystem-cellar/language-fr/format-PDF.

[9] Projet EthicAA. Livre Blanc - Éthique et agents autonomes. Projet ANR-13-CORD-0006, 2018. https://ethicaa.greyc.fr/media/files/ethicaa.white.paper.pdf.

[10] Commission européenne. Livre Blanc : Intelligence artificielle - Une approche européenne axée sur l'excellence et la confiance, 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf.

[11] J. Fjeld, N. Achten, H. Hilligoss, A.C. Nagy, and M. Srikumar. Principled Artificial Intelligence : Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Technical report, The Berkman Klein Center for Internet & Society, Harvard University, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482.

[12] E. Gibney. The battle to embed ethics in AI research. *Nature*, 577 :609, 2020. https://media.nature.com/original/magazine-assets/d41586-020-00160-y/d41586-020-00160-y.pdf.

[13] M. Hunyadi. *La Tyrannie des modes de vie – Sur le paradoxe moral de notre temps*. Le Bord de l'eau, 2015.

[14] Stanford Human Centered Artificial Intelligence. Artificial Intelligence Index - 2019 Annual Report. Stanford University , 2019. https://hai.stanford.edu/research/ai-index-2019.

[15] Journal Officiel. Loi 2019-1428 d'orientation des mobilités, chapitre II, section 1 « Véhicules autonomes et véhicules connectés », 24 décembre 2019. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000039666574.

[16] Laboratoire national de métrologie et d'essais (LNE). Évaluer les intelligences artificielles, 2021. https://www.lne.fr/fr/on-en-parle/evaluer-intelligence-artificielle-ia.

[17] B. Levin and P. Kunz. ERCIM Workshop on Quality of AI. *ERCIM News*, 123 :4–5, 2020. https://ercim-news.ercim.eu/en123/joint-ercim-actions/ercim-workshop-on-quality-in-ai.

[18] P. Marks. Can the Biases in Facial Recognition Be Fixed ; Also, Should They ? *Communications of the ACM*, 64 :3 :20–22, 2021. https://cacm.acm.org/magazines/2021/3/250698-can-the-biases-in-facial-recognition-be-fixed-also-should-they/fulltext.

[19] B. Mittelstadt. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1 :501–507, 2019. https://doi.org/10.1038/s42256-019-0114-4.

[20] J. Mokander and L. Floridi. Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines*, 2021. https://doi.org/10.1007/s11023-021-09557-8.

[21] OCDE. Recommandation du Conseil sur l'intelligence artificielle OECD/LEGAL/0449, 2019. https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449.

[22] OECD.AI. Countries & initiatives overview, 2020. https://www.oecd.ai/countries-and-initiatives.

[23] Independent High-Level Expert Group on Artificial Intelligence. Policy and Investment Recommendations for Trustworthy AI. Commission européenne, 2019. https://ec.europa.eu/digital-single-market/en/news/policy-and-in

vestment-recommendations-trustworthy-artificial-intelligence.

[24] Independent High-Level Expert Group on Artificial Intelligence. The Assessment List for Trustworthy Artificial Intelligence (ALTAI). Commision européenne, 2020. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[25] World Health Organization. Measurement of and target-setting for well-being : an initiative by the WHO Regional Office for Europe, 2012. https://www.euro.who.int/__data/assets/pdf_file/0009/181449/e96732.pdf.

[26] Parlement européen. Intelligence artificielle : questions relatives à l'interprétation et à l'application du droit international. Résolution du 9 janvier 2021. https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_FR.html.

[27] European Parliament. Framework of ethical aspects of artificial intelligence, robotics and related technologies, 2020. https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.html.

[28] T.M. Powers and J.-G. Ganascia. The Ethics of the Ethics of AI. In M.D. Dubber, F. Pasquale, and S. Das, editors, *The Oxford Handbook of Ethics of AI*. Oxford University Press, 2020.