Quelle éthique de l'éthique de l'intelligence artificielle ?

Alain Loute*

Which ethics for the ethics of artificial intelligence?

Contemporary development of artificial intelligence (AI) technologies is not without ethics. In fact, we are witnessing an increase in ethical charters and initiatives. However, this completely leaves in the shadows the question of the implicit background to our ethical reflection: the temporal and spatial framing of ethics. The article begins by highlighting the fact that certain individuals are attempting to impose a vision of the future, with some authors even speaking of the "colonization of the future". Second, it emphasizes the fact that ethical discussions on AI are frequently influenced by narratives that impose a scenography on reflection, namely the relationship between man and machine, which is localized and extracted from an organizational and social context. This temporal and spatial framing shapes the way we ask ethical questions and focuses our attention on certain elements of the situation. At the same time, it constitutes an "off-field". This article invites us to develop an "ethics of attention": Beyond the question of which ethical principles to mobilize, what do we pay attention to when considering the ethics of AI?

L'éthique n'est pas absente du développement contemporain des technologies d'intelligence artificielle. C'est même ni plus, ni moins à une « inflation » de chartes et initiatives éthiques que l'on assiste. Elle laisse cependant totalement dans l'ombre la question de l'arrière-plan implicite de notre réflexion éthique : le cadrage temporel et spatial de l'éthique. Dans un premier temps, l'article met en avant le fait que des acteurs cherchent à imposer une vision du futur, certains auteurs allant jusqu'à parler de « colonisation du futur ». Dans un second temps, il met à jour le fait que les discussions éthiques sur l'IA sont souvent façonnées par des récits qui imposent une scénographie à la réflexion, à savoir le rapport de l'homme face à la machine, une scène localisée et extraite d'un contexte organisationnel et social. Ce cadrage temporel et spatial façonne notre manière de poser les questions éthiques, il focalise notre attention sur certains éléments de la situation, tout en constituant un « hors-champ ». L'apport de cet article est une invitation à développer une « éthique de l'attention » : au-delà de la question des principes éthiques à mobiliser, à quoi portons-nous attention lorsque nous réfléchissons à l'éthique de l'IA ?

AFFILIATION

* Professeur à la Faculté de médecine et de médecine dentaire, UCLouvain

Force est de reconnaître que l'éthique, à tout le moins comme discours, n'est pas absente du développement contemporain des technologies d'intelligence artificielle. C'est même ni plus, ni moins à une « inflation » (1) de chartes et initiatives éthiques que l'on assiste. Leur nombre est tel que des études ont été produites afin de recenser les déclarations en la matière. A titre d'exemple, l'étude de Anna Jobin *et.* (2), souvent citée et publiée en 2019, a ainsi dénombré 84 documents contenant des principes ou guidelines éthiques relatives à l'IA. De plus, bien que ces productions soient le fait d'acteurs très divers, l'étude de Jobin *et al.* relève que tous les documents étudiés convergent autour des principes suivants : transparence, justice et équité, non-malfaisance, responsabilité et *privacy*.

Au vu d'une production massive de l'éthique et de l'apparente convergence autour de principes communs, ne peut-on pas considérer, pour reprendre l'expression de Thomas Powers et Jean-Gabriel Ganascia que l'éthique de l'IA est « un appel auquel il a été répondu » (3) ? Cependant, toujours selon ces deux auteurs, une telle réponse à l'appel de l'éthique n'est pas suffisante. Pour eux, il faut aborder de manière critique ce sursaut éthique : il faut développer une « éthique de l'éthique de l'IA ». La raison en est que les éthiciens auraient utilisé les « outils conceptuels qu'ils ont sous la main », alors que l'IA déstabiliserait fondamentalement les concepts fondamentaux de l'éthique. Ils pointent notamment les ambiguïtés conceptuelles dont souffrent les éthiques de l'IA lorsqu'elles utilisent les

concepts d'agent, d'autonomie ou d'intelligence. Ceux-ci sont souvent utilisés de manière impropre lorsqu'ils sont appliqués aux objets techniques. D'autres critiques sont plus virulentes encore et mettent en avant les risques de cette offre éthique inflationniste. Pour Luciano Floridi, un des risques de la prolifération de principes éthiques de l'IA est de constituer un « marché des principes », dans lequel les parties prenantes pourraient être tentées de « faire du shopping » et rechercher le type d'éthique le plus adapté pour justifier leurs comportements actuels (4).

Je voudrais m'inscrire dans l'ensemble de ces travaux qui visent à questionner cette offre massive d'éthique. Ma démarche ne visera pas à développer une réflexion méta-éthique, au sens d'une redéfinition des concepts de l'éthique, ni même à dénoncer les risques d'instrumentalisation de l'éthique. Elle visera plutôt à révéler un angle mort, un hors-champ de la réflexion, à savoir le « cadrage implicite » de l'éthique de l'IA. Que faut-il entendre par ce concept ? Dans le domaine de la sociologie de l'action collective, la *frame analysis* a mis en avant le travail de « cadrage » effectué par les acteurs. Le concept de cadre, inspiré du travail de Goffman, renvoie à un schème d'interprétation partagé qui donne du sens à des événements et organise l'expérience. Il permet aux individus de « localiser, percevoir, identifier et étiqueter » des situations au cours de leur vie et dans le monde en général.

J'utilise ici librement ce terme sociologique pour mettre en lumière le fait que les débats éthiques sur l'IA, au-delà des arguments mobilisés, ont souvent pour effet d'imposer un « cadrage temporel et spatial » de la réflexion qui est rarement questionné. Dans un premier temps, je mettrai ainsi en avant le fait que des acteurs cherchent à imposer une vision du futur, certains auteurs allant jusqu'à parler de « colonisation du futur ». Dans un second temps, je montrerai que les discussions éthiques sur l'IA sont souvent façonnées par des récits qui imposent une scénographie à la réflexion, à savoir le rapport de l'homme face à la machine, une scène localisée et extraite d'un contexte organisationnel et social. Ce cadrage temporel et spatial de la réflexion éthique de l'IA n'est pas sans effet. Il façonne notre manière de poser les questions éthiques, il focalise notre attention sur certains éléments de la situation, tout en constituant un « hors-champ ». L'apport de ce petit article est une invitation à développer une « éthique de l'attention » (5) : au-delà de la question des principes éthiques à mobiliser, à quoi portons-nous attention lorsque nous réfléchissons à l'éthique de l'IA ?

QUESTIONNER LE CADRAGE TEMPOREL DE L'ÉTHIQUE DE L'IA

En sus de l'insistance sur des principes généraux énoncés pour encadrer l'IA, force est de constater que de nombreux rapports éthiques sont traversés par un impératif d'anticipation, faisant de l'éthique de l'IA une « éthique du futur ». S'y joue une nouvelle déclinaison de la responsabilité, comme l'atteste la notion d'« innovation responsable » promue par la Commission Européenne. Pour Richard Owen, une telle démarche d'innovation responsable est « anticipatoire », dans le sens où elle suppose d'identifier des « impacts anticipés ou potentiellement anticipés qui peuvent émerger, qu'ils soient économiques, sociaux, environnementaux ou autres » (6). Le rapport Villani (7) « *Donner un sens à l'intelligence artificielle* » pose également un impératif d'anticipation :

> « la loi ne peut pas tout, entre autres car le temps du droit est bien plus long que celui du code. Il est donc essentiel que les « architectes » de la société numérique – chercheurs, ingénieurs et développeurs – qui conçoivent et commercialisent ces technologies prennent leur juste part dans cette mission en agissant de manière responsable. Cela implique qu'ils soient pleinement conscients des possibles effets négatifs de leurs technologies sur la société et qu'ils œuvrent activement à ces limiter ».

Le mouvement est double : anticiper *l'aval* du développement technologique, pour – *en amont* du développement technologique – « ouvrir la boîte noire technique » et modifier le design technologique afin d'empêcher les impacts éthiques négatifs. À l'appui de cette démarche, le rapport Villani imagine obliger les développeurs d'IA à réaliser un *discrimination impact assessment* afin de « les obliger à se poser les bonnes questions au bon moment » (7).

Si l'on ne peut que souligner l'importance d'anticiper les enjeux éthiques de demain, il est également essentiel d'expliciter la manière dont de nombreux acteurs cherchent à façonner notre rapport au futur. Ma conviction est que celui-ci est objet d'un intense travail de « cadrage ». En nous réappropriant un concept de Didier Bigo, nous pouvons dire que le futur est « colonisé » par des acteurs qui cherchent à imposer leur vision du futur comme matrice commune de toute anticipation du futur. Initialement, il a forgé cette expression dans le cadre d'une recherche sur les technologies de surveillance, pour désigner les prétentions et stratégies des experts qui appréhendent le futur comme un « futur antérieur, comme un futur déjà fixé, un futur dont ils connaissent les événements » (8). Ce futur commande alors ni plus ni moins le présent : la mise en avant d'un *worst-case scenario* incite ainsi à investir massivement dans ces technologies.

Il nous semble que dans le domaine de l'intelligence artificielle une dynamique de colonisation du futur s'est également mise en place. Elle prend d'abord la forme d'une inflation de « promesses technoscientifique ». Le développement technologique s'accompagne de l'intensification de promesses d'applications futures. Cette prolifération de promesses technoscientifiques doit être mise en regard de

l'évolution du marché de la connaissance et de l'évolution des modes de financement de celle-ci. « Les promesses sont ainsi des stratégies pour capter des ressources attribuées sur une base compétitive » (9). Cette économie de promesses façonne l'arrière-plan des réflexions sur l'IA. Expliciter et prendre conscience de cet arrière-plan est essentiel. Le risque sinon est de développer une éthique au futur qui, sans s'en rendre compte, reproduit la matrice du futur que constituent ces promesses. Sur le terrain, force est de constater que les promesses tardent parfois à se réaliser (10).

Un autre exemple de cadrage temporel implicite peut être illustré par les propos de Maggie De Block, ancienne ministre fédérale, affirmant que « l'e-santé a démarré comme un TGV. On n'arrêtera plus l'utilisation des technologies numériques dans le cadre des soins de santé ». De tels propos façonnent notre rapport au futur. Ils induisent tout d'abord, à travers l'image de rails rectilignes, l'idée d'un développement technologique unilinéaire. Or, pour le philosophe des techniques Andrew Feenberg, le regard historique sur le développement technologique passé permet de prendre conscience que « il y a toujours d'autres alternatives techniques viables qui auraient pu être développées à la place de celles qui ont été choisies » (11). Pour Feenberg, les fonctions techniques ne sont pas prédéterminées, « on les découvre au cours de leur développement et de leur utilisation » (11).

De plus, cette métaphore utilisée par Maggie De Block introduit l'idée d'une vitesse et d'un mouvement irrépressible. Ne pas prendre en compte et s'aligner sur celui-ci nous imposerait de rester à quai. On trouve un cadre temporel implicite dans un avis du Comité consultatif national d'éthique français. Dans son avis 129, le CCNE soutient ainsi que « l'insuffisance du recours au numérique (…) induit, sur une large échelle, des situations non éthiques au sein de notre système de santé ». Cette avis constitue la contribution du CCNE à la révision de la loi de bioéthique. A travers les propos rapportés ci-dessus, il s'agit pour le comité de mettre en garde : au vu des bénéfices escomptés du numérique en santé, une trop forte régulation, « une approche trop fermée de ces questions aboutirait à laisser la France à l'écart d'un mouvement inédit d'innovation ».

S'il ne faut aucun doute que le numérique constitue un élément de réponse à de nombreux défis de nos systèmes de santé, une telle conception de l'innovation comme un mouvement irrépressible ne risque-t-il pas de focaliser notre attention sur le numérique comme seul réponse aux défis de demain ? La vitesse prétendue de l'innovation ne risque-t-elle pas d'occulter que certains impacts du développement technologique se donne à voir parfois sur des temporalités plus longues ? Le philosophie Xavier Guchet nous rappelle que « Bergson faisait remarquer à propos de la machine à vapeur que plus d'un siècle après son invention, l'on commençait à peine à en ressentir la puissance de transformation – non seulement sur le travail et la production, mais aussi sur les idées » (12).

QUESTIONNER LE CADRAGE SPATIAL DE L'ÉTHIQUE DE L'IA

Dans le second temps de cet article, je voudrais expliciter le cadre spatial implicite de nombreuses réflexions éthiques sur l'IA. De manière subreptice, des récits peuvent focaliser notre attention sur une scène locale, celle du face à face de l'homme et de la machine. De telles scènes ont été médiatisées, comme par exemple l'affrontement en 2016 au jeu de go entre Lee Sedol et un programme conçu par Google Deepmind. AlphaGo a gagné toutes les parties sauf la quatrième. Ma conviction est que la médiatisation de ces récits n'est pas sans performativité : elle a pour effet d'abstraire la réflexion sur l'IA de son contexte. Elle focalise notre attention sur deux personnages du récit que sont les deux joueurs de go, à savoir Lee Sedol et Deepmind. Elle les abstrait de tout un réseau d'acteurs et d'objets : les concepteurs du programme, les organisateurs de la rencontre, le public, etc., en résumé le milieu socio-technique dans lequel ils s'insèrent.

Un autre exemple d'un cadrage spatial peut être trouvé dans l'éthique des machines proposée par le projet *Moral Machine* (https://www.moralmachine.net). Cette plate-forme met en scène des dilemmes moraux impliquant des machines comme les voitures autonomes. A titre d'exemple, dans une situation où les freins d'une voiture autonome ne fonctionneraient plus, quelle option celle-ci devrait suivre : s'écraser sur un mur et dès lors tuer son passager ou poursuivre sa route et tuer un piéton ? L'objectif de ce projet est double : d'une part, constituer une forme d'expérimentation sociale afin de « construire une image de l'opinion humaine sur la façon avec laquelle les machines doivent prendre des décisions lorsqu'elles sont confrontées à des dilemmes moraux », d'autre part, sensibiliser les médias et l'opinion publique au sujet de l'éthique des machines.

Ces récits et ces expérimentations éthiques peuvent paraitre assez anodins. Ils ont néanmoins un effet qu'une éthique de l'attention doit questionner : ils induisent l'image de l'intelligence artificielle comme un ensemble d'entités discrètes, autonomes et isolables d'un milieu socio-technique. Au contraire de cette image, de nombreux philosophes attirent notre attention sur le fait que les technologies transforment les milieux dans lesquels elles s'insèrent. Luciano Floridi défend ainsi l'idée que l'utilisation de l'IA implique ce qu'il appelle un « enveloppement », c'est-à-dire l'adaptation de l'environnement et des tâches que l'on cherche à autonomiser aux

capacités de l'IA. Pour lui, « nous ne construisons pas des véhicules autonomes en plaçant des androïdes à la place du conducteur, mais en repensant l'ensemble de l'écosystème des véhicules et des environnements » (4). La thèse de Floridi est forte : c'est l'environnement qui est adapté à l'IA et non l'inverse.

Tout une littérature sociologique a également mis en lumière le travail humain qu'occulte la focalisation médiatique sur la confrontation de la machine et de l'homme. Pour un auteur comme Antonio Casilli, ces récits nous font oublier que l'intelligence artificielle est artificielle, c'est-à-dire qu'elle est produite par ce qu'il appelle le « digital labor » : un travail de production des données, d'accompagnement des machines dans leur apprentissage, etc. La panne d'un dispositif technique peut également être l'occasion de prendre conscience tant de l'infrastructure matérielle invisibilisée, que des interdépendances et du travail humain impliqués dans le fonctionnement d'un objet technique.

Je voudrais terminer cette seconde partir en illustrant mon propos à partir d'une recherche menée en collaboration avec le sociologue Gérald Gaglio (13). Nous avons mené un travail d'enquête de terrain portant sur la manière dont des logiciels d'IA sont appropriés par les radiologues, en sénologie et traumatologie. Il s'agissait de dispositifs d'aide à la décision qui se prononcent sur des images en désignant des zones potentiellement pathologiques et/ou en pronostiquant un score global de risque (de cancer par exemple).

L'examen de plusieurs cas d'usage nous a convaincus de la nécessité de nous questionner sur la délimitation de notre objet de recherche. Alors que de prime abord notre objet de recherche était l'appropriation de l'outil par les radiologues, il nous a semblé que nous assistions davantage à l'expérimentation de nouveaux modes d'organisation du travail par le biais de l'utilisation des logiciels d'IA. L'innovation était avant tout organisationnelle et gestionnaire, l'IA nous semblant avant tout mobilisée pour transformer un milieu socio-technique. Un des cas étudiés concerne un cabinet de radiologie situé au centre-ville d'une grande ville du sud-est de la France. Ce cabinet a été racheté par un groupe financier deux ans auparavant, suite à un regroupement, trois ans plus tôt, avec plusieurs autres cabinets. Deux logiciels d'IA (en sénologie et en traumatologie) ont été acquis par le groupe financier, sans consultation du radiologue-gérant ou de ses équipes. Il s'agit de logiciels qui s'appuient sur des algorithmes de *deep learning* entraînés selon la méthode de l'apprentissage supervisé.

Lors d'une première visite, le radiologue-gérant nous a précisé qu'il analyse toujours les images, avant de lire l'avis du logiciel de détection. Il s'agit en quelque chose d'éviter un « biais de confirmation » et d'éviter de ne porter son attention que là où le logiciel l'attire. Lors d'une deuxième visite, la surcharge de travail du radiologue-gérant nous a surpris. La manipulatrice nous indique que lors de la semaine, il a dû en moyenne interpréter plus de 200 examens par jour. Il doit aussi se prononcer (par le biais d'un dispositif de téléradiologie) sur des examens issus d'un autre cabinet qui vient d'ouvrir dans le département et qui appartient au même groupe. Contrairement à ce que le radiologue indiquait en entretien la première fois, il regarde d'abord l'affichage du logiciel indiquant « fract » « no fract » ou « doubt », puis se penche sur les images ensuite. Cet enquêté a tellement d'examens à interpréter qu'il va au plus vite, et la confiance qu'il attribue à la « valeur prédictive négative » y aide : si le logiciel ne suspecte rien, il passe rapidement à un autre examen.

Ce cas d'étude montre l'importance de prendre en compte le contexte dans lequel s'insère les usages. En effet, les logiciels de détection arrivent dans un contexte de rachat massif de cabinets de radiologues-propriétaires qui s'approchent de la retraite et ne trouvent pas de repreneurs, auxquels s'ajoutent le manque de radiologues sur de nombreux territoires et l'augmentation du nombre d'examens prescrits. Ce contexte va nécessairement conditionner la manière dont l'IA va s'implémenter. De plus, on peut également se demander si l'objectif du groupe financier n'est pas avant tout gestionnaire, l'IA permettant d'augmenter la productivité des radiologues.

CONCLUSION

L'offre éthique en matière d'intelligence artificielle a pris principalement la forme d'une « éthique à principes » (14) qui énonce des préceptes généraux (respect de l'autonomie, bienfaisance, etc.) afin d'encadrer la mise en application de l'IA. Cette éthique à principes à la mérite d'avoir mis sur l'agenda le thème de l'éthique de l'IA et constitue des repères pour justifier certains usages et l'IA et en dénoncer d'autres. Elle laisse cependant totalement dans l'ombre la question du cadrage implicite de notre réflexion.

Pour expliciter ce cadrage, il nous semble tout d'abord essentiel de nous rendre attentif au contexte d'inflation de promesses technoscientifiques. Il nous semble important également de questionner les visions unilinéaires du développement technologique et l'injonction à innover. Sur le plan du cadrage spatial de notre réflexion, il faut pouvoir questionner la focalisation sur certains scènes et élargir les échelles de la réflexion. Interroger les contours des objets de notre réflexion, être attentif à la manière dont les technologies constituent et transforment nos relations et milieux. Ce texte s'est voulu une invitation à interroger ce cadrage implicite. Réfléchir à ce dont nous portons attention est une tâche à laquelle l'éthique de l'IA devrait s'affronter.

RÉFÉRENCES

1. Benbouzid B, Cardon D. Contrôler les IA. Réseaux 2022; 232-233; 9-26. doi.org/10.3917/res.232.0009.
2. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019; 1; 389-399. doi.org/10.1038/s42256-019-0088-2.
3. Powers T, Ganascia JG. The Ethics of the Ethics of AI. In: Dubber MD, Pasquale F, Das S. editors. The Oxford Handbook of Ethics of AI. New York: Oxford University Press; 2020; 26-51.
4. Floridi L. L'éthique de l'intelligence artificielle: Principes, défis et opportunités. Sesto San Giovanni: Mimésis; 2023.
5. Grandjean N, Loute A. Valeurs de l'attention : Perspectives éthiques, politiques et épistémologiques. Lille: Septentrion; 2019.
6. Owen R. Des valeurs coopératives à l'innovation responsable. Projectics / Proyéctica / Projectique 2012; 11-12; 13-21.
7. Rapport Villani, Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. 8 mars 2018.
8. Bigo B. Sécurité maximale et prévention ? La matrice du futur antérieur et ses grilles. in B. Cassin (éd.), *Derrière les grilles, Sortons du tout-évaluation*, Paris, Fayard, 2014, p. 111-138, p. 126.
9. Audetat M. Introduction: Sciences et technologies émergentes : pourquoi tant de promesses ? in M. Audétat (éd.), *Sciences et technologies émergentes : pourquoi tant de promesses ?*, *op. cit.*, p. 5-27, p. 11.
10. Van Hout A., Willems D, Hettinga M, Pols J, Pourquoi les attentes suscitées par la télésurveillance sont souvent déçues: Étude ethnographique d'un dispositif de télésuivi infirmier en soins palliatifs. Réseaux 2018; 207; 95-121. doi.org/10.3917/res.207.0095.
11. Feenberg A, (Re)penser la technique: Vers une technologie démocratique. Paris: La Découverte; 2004.
12. Guchet X. Philosophie des nanotechnologies. Paris: Hermann; 2014.
13. Gaglio G., Loute, A. L'émergence d'enjeux éthiques lors d'expérimentations de logiciels d'intelligence artificielle: Le cas de la radiologie. Réseaux 2023; 240(4); 145-178. doi.org/10.3917/res.240.0145.
14. Ganascia J.G. Servitudes virtuelles. Paris: Seuil; 2022.