Luciano Floridi

L'ÉTHIQUE DE L'INTELLIGENCE ARTIFICIELLE

Principes, défis et opportunités

Préface de Jean-Gabriel Ganascia

Édition française éditée et traduite par
Enrico Panaï et Emmanuel R. Goffi

ÉDITIONS MIMÉSIS

www.editionsmimesis.fr
e-mail : info@editionsmimesis.fr
Collection : *Philosophie*, n. 93
ISBN : 9788869763762

P.I. C.F. 02419370305
Cedif Diffusion
Pollen Distribution

SOMMAIRE

PRÉFACE DE L'AUTEUR

L'éducation, le commerce et l'industrie, les voyages et la logistique, la banque, la vente au détail et le shopping, le divertissement, le bien-être et les soins de santé, la politique et les relations sociales – bref, la vie elle-même telle que nous la connaissons aujourd'hui – sont tous devenus inconcevables sans la présence des technologies, services, produits et pratiques numériques. Quiconque n'est pas perplexe devant une telle révolution numérique, n'en a pas saisi l'ampleur. Nous parlons d'un nouveau chapitre de l'histoire de l'humanité. Bien sûr, de nombreux autres chapitres ont eu lieu auparavant. Ils ont tous eu la même importance. L'humanité a connu un monde avant et après la roue, le travail du fer, l'alphabet, l'imprimerie, le moteur, l'électricité, les automobiles, la télévision et le téléphone. Chaque transformation était unique. Certaines ont changé de manière irréversible notre perception de nous-mêmes, de la réalité et de l'expérience que nous en avons, entraînant des conséquences complexes et à long terme. Nous trouvons encore, par exemple, de nouvelles façons d'exploiter la roue (il suffit de penser à la roue à clic de l'iPod). De même, ce que l'humanité va réaliser grâce aux technologies numériques est inimaginable. Comme je le souligne au Chapitre 1, personne en 1964 n'aurait pu deviner ce que serait le monde seulement cinquante ans plus tard. Les futurologues sont les nouveaux astrologues : nous ne devrions pas nous fier à eux. Pourtant, il est également vrai que la révolution numérique ne se produira qu'une seule fois, et elle se produit maintenant. Une page de l'histoire de l'humanité a été tournée et un nouveau chapitre a commencé. Les générations futures ne sauront jamais à quoi ressemblait une réalité exclusivement analogique, hors ligne et pré-numérique. Nous sommes la dernière génération à l'avoir vécu.

Le prix à payer pour une place si particulière dans l'histoire est celui d'incertitudes inquiétantes. Les transformations induites par les technologies numériques sont époustouflantes. Elles justifient une certaine confusion et une certaine appréhension : il suffit de regarder les titres des journaux. Toutefois, la place particulière que nous occupons dans ce tournant historique, entre une réalité entièrement analogique et une réalité de plus en plus numérique, est également porteuse d'opportunités extraordinaires. C'est précisément parce que la révolution numérique vient de commencer que nous avons la possibilité de la façonner de manière positive, au bénéfice de l'humanité et de notre planète. Comme l'a dit un jour Winston Churchill, « Nous façonnons nos bâtiments ; ce sont eux qui nous façonnent ensuite ». Nous n'en sommes qu'au tout début de la construction de nos réalités numériques. Nous pouvons les construire correctement, avant qu'elles ne commencent à nous influencer et à influencer les générations futures de manière erronée. Il ne s'agit pas d'être pessimiste ou optimiste. Il est inutile de se demander si le verre est à moitié vide ou à moitié plein. La question intéressante est de savoir comment nous pouvons le remplir. Cela signifie qu'il faut s'engager de manière constructive dans l'analyse éthique des problèmes et dans la conception des bonnes solutions.

Pour identifier la meilleure voie à suivre pour le développement de nos technologies numériques, la première étape essentielle est de mieux comprendre. Nous ne devons pas nous lancer dans la création d'un monde de plus en plus numérique en somnambule. L'insomnie de la raison est vitale car son sommeil génère des erreurs monstrueuses, parfois irréversibles. Il est essentiel de comprendre les transformations technologiques qui se déroulent sous nos yeux si nous voulons orienter la révolution numérique dans une direction qui soit à la fois socialement préférable (équitable) et écologiquement durable. Cela ne peut être qu'un effort de collaboration (Floridi à paraître). Ainsi, dans ce livre, je propose ma contribution en partageant quelques idées sur un type de technologie numérique, à savoir l'intelligence artificielle (IA), et sur la question spécifique de son éthique.

Ce livre fait partie d'un projet de recherche plus vaste sur les transformations de l'agentivité (la *capacité d'agir* ou la capacité

d'interagir avec le monde et d'en tirer des enseignements en vue d'atteindre un objectif) induites par la révolution numérique. Au départ, je pensais pouvoir travailler à la fois sur l'intelligence artificielle – entendue comme *forme d'« agir artificiel* », le sujet de ce livre – et sur l'agir politique – entendue comme *forme d'« agir collectif* » soutenue et influencée par les interactions numériques. Lorsque j'ai été invité à donner les *conférences Ryle* (Ryle Lectures) en 2018, c'est précisément ce que j'ai tenté de faire. J'ai présenté les deux sujets comme deux aspects d'une même transformation, plus profonde. Les organisateurs et les participants m'ont dit (peut-être par gentillesse) que je n'avais pas failli. Mais personnellement, je n'ai pas trouvé que c'était un grand succès. Ce n'est pas parce que l'approche de l'éthique de l'IA et de la politique de l'information à partir d'un point de vue unique, fondé sur la capacité d'agir, ne fonctionne pas, mais parce que cela ne fonctionne bien que si l'on est prêt à passer sur les détails et à échanger la portée contre la profondeur. Cela peut convenir dans le cadre d'une série de conférences, mais le fait de couvrir les deux sujets dans une seule monographie de recherche aurait donné un ouvrage encore moins attrayant que ce livre. C'est pourquoi, suivant les conseils avisés de Peter Momtchiloff, d'Oxford University Press, j'ai décidé de scinder le projet en deux : ce livre sur l'éthique de l'IA et un second livre sur la politique de l'information. C'est le bon endroit pour indiquer au lecteur où trouver les deux livres dans le cadre du projet global.

Ce livre est la première partie du quatrième volume d'une tétralogie qui comprend *La philosophie de l'information* (volume I, Floridi, 2011), *L'éthique de l'information* (volume II, Floridi, 2013), et *La logique de l'information* (volume III, Floridi, 2019d) [éd. pas encore traduite en français]. J'ai intitulé cette tétralogie *Principia Philosophiae Informationis* non pas comme un signe d'orgueil démesuré (bien que cela puisse l'être), mais comme un jeu de mots interne entre certains collègues. Dans une sorte de compétition d'aviron, j'ai plaisanté en disant qu'il était temps pour Oxford de rattraper Cambridge sur le score de 3-0 pour « *principia* ». Ce n'était pas un jeu de mots que beaucoup ont trouvé drôle ou même intelligible.

Dans le cadre du projet *Principia*, ce livre occupe une position intermédiaire entre le premier et le deuxième volume (un peu comme

le troisième volume), car l'épistémologie, l'ontologie, la logique et l'éthique contribuent toutes au développement des thèses présentées dans les chapitres suivants. Mais comme le lecteur est en droit de s'y attendre, tous les volumes sont écrits de manière autonome, de sorte que ce livre peut être lu sans aucune connaissance de ce que j'ai pu publier par ailleurs. Néanmoins, les volumes sont complémentaires. Le message essentiel du premier volume est relativement simple : l'information sémantique est une donnée bien structuré, significative et fiable ; la connaissance est une information sémantique pertinente correctement prise en compte ; les humains sont les seuls moteurs sémantiques et organismes informationnels conscients connus qui peuvent concevoir et comprendre des artefacts sémantiques, et ainsi développer une connaissance croissante de la réalité et d'eux-mêmes, en tant que capital sémantique ; et la réalité est mieux comprise comme la totalité de l'information (remarquez l'absence cruciale de « sémantique »).

Dans ce contexte, le deuxième volume étudie les fondements de l'éthique des organismes informationnels (*infosrgs*) comme nous, qui s'épanouissent dans des environnements informationnels (l'*infosphère*[1]) et sont responsables de leur construction et de leur bien-être. En bref, le deuxième volume traite de l'éthique des inforgs de l'infosphère qui font de plus en plus l'expérience de la vie en tant que « *onlife* » (Floridi, 2014b) – à la fois *online* et *offline*, analogique et numérique. Dans une démarche kantienne classique, nous passons ainsi de la philosophie théorique à la philosophie pratique (au sens de *praktischen*, et non au sens d'utile ou d'appliqué). Le troisième volume se concentre sur la logique conceptuelle de l'information sémantique en tant que *modèle*. Il est lié à l'analyse épistémologique fournie dans *La philosophie de l'information*. Dans la mesure où

1 L'« infosphère » est un mot que j'ai détourné il y a plusieurs années pour désigner l'ensemble de l'environnement informationnel constitué par toutes les entités informationnelles (y compris les agents informationnels), leurs propriétés, leurs interactions, leurs processus et leurs relations mutuelles. Il s'agit d'un environnement comparable au cyber-espace (qui n'est qu'une de ses sous-régions), mais également différent, car il comprend également des espaces d'information hors ligne et analogiques. C'est un environnement, et donc un concept, qui évolue rapidement. Voir https://fr.wikipedia.org/wiki/Infosph%C3%A8re

le volume se concentre sur la logique conceptuelle de l'information sémantique en tant que *modèle*, il offre un pont vers l'analyse normative fournie dans *L'éthique de l'information*. Le troisième volume traite, entre autres, des devoirs, des droits et des responsabilités associés aux pratiques poïétiques qui caractérisent notre existence, qu'il s'agisse de donner un sens au monde ou de le modifier en fonction de ce que nous considérons comme moralement bon et normativement juste. Fonctionnant comme une charnière entre les deux livres précédents, le troisième volume prépare la base du quatrième volume sur la *Politique de l'information*, dont ce livre constitue la première partie. Ici, le constructionnisme épistémologique, normatif et conceptuel développé dans les volumes précédents soutient l'étude des possibilités de conception qui s'offrent à nous. Ces possibilités nous permettent de comprendre et de façonner ce que j'appelle « le projet humain » dans nos sociétés de l'information, en concevant correctement de nouvelles formes d'agence artificielle et politique. Les principales thèses de ce livre sont que l'IA est rendue possible par le découplage de la capacité d'agir et de l'intelligence, et qu'il est donc préférable de comprendre l'IA comme une nouvelle forme d'agir, et non d'intelligence ; que l'IA est donc une révolution étonnante, mais dans un sens pragmatique et non cognitif, et que les défis et les opportunités concrets et urgents concernant l'IA émergent du fossé entre l'agir et l'intelligence, qui continuera de s'élargir à mesure que l'IA deviendra de plus en plus performante. Dans l'ensemble, on peut considérer que les quatre volumes cherchent à inverser ce que je crois être plusieurs idées fausses. Ces idées fausses s'expliquent facilement à l'aide du modèle de communication classique, introduit par Shannon, qui comprend l'émetteur, le message, le récepteur et le canal (Shannon et Weaver, 1975 ; 1998). L'épistémologie se concentre trop sur le récepteur/consommateur passif de la connaissance alors qu'elle devrait s'intéresser à l'émetteur/producteur actif. Elle devrait passer de la *mimesis* à la *poiesis*. C'est parce que connaître, c'est concevoir. L'éthique se concentre trop sur l'émetteur/agent, alors qu'elle devrait s'intéresser au récepteur/patient et, surtout, à la relation entre l'émetteur et le récepteur. En effet, l'attention, le respect et la tolérance sont les clés du bien. La métaphysique se concentre trop sur les relations, émetteur/producteur/

agent/récepteur/consommateur/patient (qu'elle conçoit comme des entités), alors qu'elle devrait s'intéresser au message/relations. En effet, les structures dynamiques constituent le structuré. La logique se concentre trop sur les canaux de communication comme support, justification ou fondement de nos conclusions, alors qu'elle devrait également s'intéresser aux canaux qui nous permettent d'extraire (et de transférer) des informations de diverses sources de manière fiable. La raison en est que la logique de la conception de l'information est une logique de relations et de groupes de relations plutôt qu'une logique de choses en tant que porteurs de prédicats. L'IA, ou du moins sa philosophie, se concentre trop sur l'ingénierie d'une forme d'intelligence de type biologique, alors qu'elle devrait s'intéresser à l'ingénierie d'artefacts qui peuvent fonctionner avec succès sans aucun besoin d'intelligence. En effet, l'IA n'est pas un mariage mais un divorce entre la capacité de résoudre un problème ou de mener à bien une tâche en vue d'un objectif et la nécessité d'être intelligent en le faisant. Comme l'illustre assez bien mon téléphone qui joue aux échecs mieux que tous ceux que je connais, l'IA est la continuation d'un comportement intelligent par d'autres moyens. Enfin, la politique (le sujet de la seconde moitié du quatrième volume) ne consiste pas à gérer notre *res publica*, mais à prendre soin des relations qui nous rendent sociaux – notre *ratio publica*. Je serais étonné qu'un seul de ces revirements dans nos paradigmes philosophiques soit couronné de succès.

Permettez-moi maintenant de donner un bref aperçu du contenu de ce livre. La tâche de ce volume est encore de contribuer, comme les précédents, à l'élaboration d'une philosophie de notre temps pour notre temps, comme je l'ai écrit plus d'une fois. Comme dans les volumes précédents, il le fait de manière systématique (l'architecture conceptuelle est considérée comme un élément précieux de la pensée philosophique), plutôt que de manière exhaustive, en poursuivant deux objectifs.

Le premier objectif est d'ordre méta-théorique et est atteint par la première partie du volume, qui comprend les trois premiers chapitres. J'y propose une interprétation du passé (Chapitre 1), du présent (Chapitre 2) et de l'avenir de l'IA (Chapitre 3). La première partie n'est ni une introduction à l'IA dans un sens technique, ni

une sorte d'IA pour débutants. Il existe déjà beaucoup d'excellents livres sur le sujet, et je recommande le classique (Russell et Norvig, 2018) à toute personne intéressée. Il s'agit plutôt d'une interprétation philosophique de l'IA en tant que technologie. Comme je l'ai déjà mentionné, la thèse centrale est que l'IA représente un divorce sans précédent entre la capacité d'agir et l'intelligence. Sur cette base, la deuxième partie de l'ouvrage poursuit une investigation non pas méta-théorique mais théorique des conséquences du divorce évoqué plus haut. C'est le deuxième objectif. Le lecteur ne doit pas s'attendre à ce qu'un manuel traite de tous les principaux problèmes éthiques liés à l'IA. De nombreux livres couvrent déjà tous les sujets pertinents de manière systématique, et je recommanderais parmi eux : Dignum (2019), Coeckelbergh (2020), Moroney (2020), Bartneck *et al*. (2021), Vieweg (2021), Ammanath (2022), Blackman (2022) ; et les manuels suivants : Dubber, Pasquale et Das (2020), DiMatteo, Poncibò et Cannarsa (2022), Voeneky *et al*. (2022). La deuxième partie développe plutôt l'idée que l'IA est une nouvelle forme d'agir qui peut être exploitée de manière éthique et non éthique. Plus précisément, au Chapitre 4, j'offre une perspective unifiée sur les nombreux principes qui ont déjà été proposés pour encadrer l'éthique de l'IA. Cela conduit à une discussion, au Chapitre 5, sur les risques potentiels qui peuvent compromettre l'application de ces principes, puis à une analyse de la relation entre les principes éthiques et les normes juridiques, ainsi qu'à la définition de l'éthique douce comme éthique post-conformité au Chapitre 6. Après ces trois chapitres, j'analyse les défis éthiques soulevés par le développement et l'utilisation de l'IA (Chapitre 7), les mauvais usages de l'IA (Chapitre 8), et les bonnes pratiques lors de l'application de l'IA (Chapitre 9). Le dernier groupe de chapitres est consacré à la conception, au développement et au déploiement de l'IA pour le bien social ou IApBS. Le Chapitre 10 traite de la nature et des caractéristiques de l'IApBS. Dans le Chapitre 11, je reconstruis les impacts positifs et négatifs de l'IA sur l'environnement et comment elle peut être une force positive dans la lutte contre le changement climatique – mais pas sans risques et coûts, qui peuvent et doivent être évités ou minimisés. Dans le Chapitre 12, je développe l'analyse présentée aux Chapitres 9 et 10 pour examiner la possibilité d'utiliser l'IA pour

soutenir les 17 objectifs de développement durable (ODD) des Nations unies (Cowls, Png et Au, 2019). J'y présente l'Oxford Initiative on AIxSDG (AIxODD dans la version en français), un projet que j'ai dirigé et achevé en 2022. Dans le Chapitre 13, je conclus en plaidant en faveur d'un nouveau mariage entre le Vert de tous nos habitats et le Bleu de toutes nos technologies numériques. Ce mariage peut soutenir et développer une meilleure société et une biosphère plus saine. Le livre se termine par quelques références à des concepts qui occuperont une place centrale dans le prochain livre, *La politique de l'information*, consacré (comme mentionné ci-dessus) à l'impact des technologies numériques sur l'agir socio-politique. Tous les chapitres sont strictement liés. J'ai donc ajouté des références internes chaque fois qu'elles pouvaient être utiles. Comme l'a fait remarquer un critique anonyme, ils pourraient être lus dans un ordre légèrement différent. Je suis d'accord.

Comme pour les volumes précédents, il s'agit également d'un livre allemand en termes de racines philosophiques. Il est écrit dans une perspective post-analytique-continentale, qui, à mon avis, est en train de s'estomper. Le lecteur attentif situera facilement cet ouvrage dans la tradition qui relie le pragmatisme (notamment Charles Sanders Peirce) à la philosophie de la technologie (notamment Herbert Simon)[2]. Contrairement au premier volume et encore plus que les volumes deux et trois, ce quatrième volume est moins néo-kantien que je ne l'attendais. Contrairement aux volumes deux et trois, il est aussi moins platonicien et cartésien. Bref, en l'écrivant, j'ai pris conscience que je sortais de l'ombre de mes trois héros philosophiques. Ce n'était pas prévu, mais c'est ce qui arrive quand on suit son propre raisonnement peu importe là où il mène. *Amici Plato, Cartesius et Kant*, *sed magis amica veritas*. Dans *L'éthique de l'information*, j'ai écrit que « certains livres écrivent leurs auteurs ». J'ai maintenant l'impression que seuls les mauvais livres sont entièrement contrôlés par leurs auteurs : ce sont des best-sellers d'aéroport.

Par rapport aux volumes précédents, la principale différence est que je suis désormais de plus en plus convaincu que la philosophie, dans ce qu'elle a de meilleur, est un *design conceptuel*. Le

2 Le lecteur intéressé par l'exploration de ces liens pourra consulter Allo (2010), Demir (2012), et Durante (2017).

design conceptuel permet de réaliser des projets utiles (comprendre le monde pour l'améliorer) et de le sémantiser (donner du sens à l'Être, tout en prenant soin du capital sémantique de l'humanité et en l'enrichissant). Tout a commencé par la prise de conscience d'une évidence, grâce à un cas concret concernant un très célèbre philosophe d'Oxford. Le véritable héritage de Locke est sa pensée politique, et non son épistémologie. Peut-être que Kant n'a pas voulu nous induire en erreur en nous faisant croire que l'épistémologie et l'ontologie sont les reines du royaume philosophique, mais c'est ainsi que j'ai été éduqué à penser la philosophie moderne. Peut-être que ni Wittgenstein ni Heidegger n'ont pensé que la logique, le langage et leurs philosophies devaient remplacer les deux reines comme leurs seuls héritiers légitimes, mais c'est aussi la façon dont j'ai été éduqué à penser la philosophie contemporaine. Quoi qu'il en soit, aujourd'hui, je ne place plus aucune de ces disciplines au centre de l'entreprise philosophique. Je me tourne plutôt vers l'éthique, la philosophie politique et la philosophie du droit. C'est la recherche, la compréhension, l'élaboration, la mise en œuvre et la négociation de ce qui est moralement bon et juste qui est au cœur de la réflexion philosophique. Tout le reste fait partie du voyage nécessaire pour atteindre ce lieu, mais ne doit pas être confondu avec le lieu lui-même. Le *fondationnalisme* philosophique (ce qui fonde quoi) est crucial, mais seulement en vue de l'*eschatologie* philosophique (ce qui conduit à quoi). Toute bonne philosophie est eschatologique.

En ce qui concerne le style et la structure de ce livre, je peux répéter ici ce que j'ai écrit dans la préface de tous les volumes précédents. Je reste douloureusement conscient qu'il ne s'agit pas d'un livre passionnant, et c'est un euphémisme, malgré mes efforts pour le rendre aussi intéressant et convivial que possible. Je reste convaincu que la recherche ésotérique (au sens technique) en philosophie est le seul moyen de développer de nouvelles idées. Mais la philosophie *exotérique* a une place déterminante. Elle est comme la pointe plus accessible et pertinente de la partie plus obscure mais nécessaire de l'iceberg sous la surface de la vie quotidienne. Le lecteur intéressé par une lecture beaucoup plus légère pourra consulter *The Fourth Revolution* : *How the Infosphere is Reshaping Human Reality* (Floridi, 2014a) ou peut-être l'ouvrage encore plus facile *Information*

– *A Very Short Introduction* (Floridi, 2010b) [éd. pas encore traduite en français].

Comme je l'ai déjà écrit, ce livre demande malheureusement non seulement de la patience et un peu de temps, mais aussi une certaine ouverture d'esprit. Ce sont des ressources rares. Au cours des trois dernières décennies de débats, on m'a fait prendre pleinement conscience – parfois de manière beaucoup moins amicale que je ne le souhaiterais – que certaines des idées défendues dans cet ouvrage, ainsi que dans les précédents, sont controversées. Elles ne sont pas censées l'être intentionnellement. Dans le même temps, j'ai également remarqué que l'on commet souvent des erreurs en se fiant aux « attracteurs systémiques » : si une nouvelle idée ressemble un peu à une ancienne idée que nous avons déjà, alors l'ancienne est un aimant vers lequel la nouvelle est puissamment attirée, presque irrésistiblement. Nous finissons par penser que « ce nouveau » est exactement comme « cet ancien » et que, par conséquent, soit « cet ancien » peut être écarté, soit, si nous n'aimons pas « cet ancien », nous n'aimons pas non plus « ce nouveau ». C'est une mauvaise philosophie en effet, mais il faut de la force mentale et de l'exercice pour résister à un changement aussi puissant. Je le sais. Je parle en tant que pécheur. Dans le cas de ce livre, je crains que certains lecteurs ne soient tentés de conclure qu'il s'agit d'un livre anti-technologie, un livre dans lequel j'indique les limites de l'IA ou ce que « l'IA ne peut pas faire ». Ils pourraient également conclure le contraire, à savoir que ce livre est trop optimiste à l'égard de la technologie, trop amoureux de la révolution numérique et de l'IA comme panacée. Ces deux conclusions sont erronées. Ce livre est une tentative de rester au milieu, dans un endroit qui n'est ni l'enfer ni le paradis, mais le purgatoire laborieux des efforts humains. Bien sûr, je serais déçu si on me disait que j'ai échoué malgré ma tentative. Mais je serais encore plus déçu et frustré si cette tentative était mal comprise. Il existe de nombreuses façons d'apprécier la technologie. L'une d'entre elles est celle de la bonne conception et de la gouvernance éthique, et je pense que c'est la meilleure approche. Le lecteur n'est pas obligé de me suivre aussi loin, mais il ne faut pas se méprendre sur la direction que je prends.

Comme dans les volumes précédents, j'ai prévu des résumés et des conclusions au début et à la fin de chaque chapitre, ainsi qu'une certaine redondance, pour aider le lecteur à accéder plus facilement au contenu de ce livre. En ce qui concerne la première caractéristique, je sais qu'elle est légèrement peu orthodoxe. Mais la solution consistant à commencer chaque chapitre par un « précédemment, au chapitre *x*… » devrait permettre au lecteur de parcourir le texte ou d'avancer rapidement dans des chapitres entiers, sans perdre l'essentiel de l'intrigue. Les amateurs de science-fiction qui reconnaîtront la référence à *Battlestar Galactica*, qui reste l'une des meilleures séries que j'aie jamais regardées, pourront considérer ce quatrième volume comme l'équivalent de la quatrième saison. J'ai essayé de convaincre un ancien correcteur de me laisser utiliser l'expression « précédemment, *au* chapitre *x*… », mais cela m'a semblé trop exagéré sur le plan linguistique. L'un des relecteurs anonymes a suggéré de supprimer le court résumé au début de chaque chapitre. J'ai décidé de les conserver, d'une part parce que je pense que, au mieux, ils sont utiles et, au pire, ils ne font pas de mal, et d'autre part parce que c'est une caractéristique de tous les livres que j'ai publiés dans le cadre de ce projet étendu.

En ce qui concerne la deuxième caractéristique, j'ai décidé, lors de l'édition de la version finale du livre, de laisser certaines répétitions et reformulations de thèmes récurrents dans les chapitres, chaque fois que je pensais que l'endroit où le contenu original avait été introduit était trop éloigné en termes de pages ou de contexte théorique. Si le lecteur a parfois une impression de *déjà-vu*, j'espère que ce sera au profit de la clarté – comme une caractéristique, pas comme un bug.

Un dernier mot maintenant sur ce que le lecteur ne trouvera pas dans les pages qui suivent. Il ne s'agit pas d'une introduction à l'IA ou à l'éthique de l'IA. Je ne cherche pas non plus à fournir une étude exhaustive de toutes les questions qui pourraient être qualifiées d'« éthique de l'IA ». Les réviseurs anonymes ont recommandé de supprimer plusieurs courts chapitres dans lesquels je cherchais à appliquer les idées développées dans la deuxième partie, et de les mettre en ligne. Le lecteur intéressé les trouvera sur SSRN, sous mon nom. J'espère travailler de manière plus approfondie sur un

audit de l'IA basé sur l'éthique et sur l'acquisition éthique de l'IA (deux sujets cruciaux qui sont encore peu explorés) et j'ai laissé les considérations plus géopolitiques sur les politiques de l'IA à *La politique de l'information*. Le lecteur intéressé pourra également consulter Cath *et al*. (2018) sur les approches des États-Unis (US), de l'Union européenne (UE) et du Royaume-Uni en matière d'IA, ou Roberts, Cowls, Morley *et al*. (2021) et Hine et Floridi (2022) pour l'approche chinoise. Ce n'est pas non plus un livre sur les aspects statistiques et informatiques des questions dites ERT (équité, responsabilité et transparence ou FAT – *fairness, accountability, and transparency* – en anglais) ou XAI (IA explicable), ni sur la législation les concernant. Ces sujets ne sont abordés que dans les chapitres suivants[3]. Il s'agit d'un livre philosophique sur certaines des racines – mais pas sur les feuilles – de certains des problèmes liés à l'IA de notre époque. Il traite d'une nouvelle forme d'agir, de sa nature, de sa portée et de ses défis. Il s'agit de savoir comment exploiter cette agence au profit de l'humanité et de l'environnement.

3 Pour plus d'informations sur ces sujets, voir Watson et Floridi (2020), Lee et Floridi (2020), Lee, Floridi et Denev (2020).