Chapitre 8

L'IA explicable - La panacée pour une IA éthique ?

Dragos Vieru *Renée-Maria Schmitt* *Élodie Boissières*

Paru dans D. Lemire, et N. Mezghani (Eds.) (2024) : *La science des données : Théorie et applications avec R et Python, Éditions TÉLUQ*

8.1 Introduction

Le domaine de l'intelligence artificielle (IA) et de l'apprentissage automatique (AA ou *Machine Learning*) a fait l'objet d'une attention particulière ces derniers temps, notamment avec l'émergence de modèles d'IA génératifs tels que ChatGPT. L'IA est un terme général qui englobe divers domaines de recherche tels que le traitement du langage naturel, la parole, la vision, la robotique et l'AA (Mukhamediev et al., 2022). Bien qu'il n'existe pas encore de définition uniforme de l'IA, il est clair qu'il s'agit d'un domaine passionnant qui évolue rapidement (Ali et al., 2023).

Adoptés dans divers domaines et contextes, tels que l'éducation, le droit, les soins de santé, la finance et les transports (Agarwal et al., 2022 ; Jia et Zhang, 2022), les systèmes d'IA deviennent omniprésents et conduisent à une transition vers une société plus algorithmique (Adadi et Berrada, 2018). Ils remodèlent nos valeurs, nos propriétés et nos interactions quotidiennes (Floridi et al., 2018 ; Zhou et al., 2022).

Cependant, il est actuellement controversé et difficile de mesurer l'influence et les impacts, positifs ou négatifs des systèmes d'IA, sur les vies et les environnements humains (Floridi et al., 2018). En effet, même si les systèmes d'IA impactent de nombreux domaines, on peut penser que les répercussions seront moindres lorsque l'on parle de recommandations musicales et plus critiques lorsque l'on parle du traitement médical ou du transport autonome (Arrieta et al., 2020). Ainsi, les principaux défis et risques qui émergent des systèmes d'IA sont principalement issus des décisions négatives ou néfastes de ces systèmes (Henriksen et al., 2021).

L'un de ces principaux défis concerne le manque de transparence des systèmes d'IA, notamment à cause de leur complexité alogrithmique. Il devient très difficile, voire

impossible, pour les humains de comprendre les systèmes d'IA opaques, leur fonctionnement interne et leur raisonnement pour la prise de décision. Or, la transparence et/ou la compréhension des systèmes d'IA sont essentielles pour garantir leur comportement et leur fonctionnement corrects, particulièrement en ce qui concerne leur alignement sur les valeurs et les principes éthiques. Par conséquent, il devient crucial de comprendre le raisonnement qui sous-tend les décisions prises par les systèmes d'IA, et les humains doivent être en mesure de recueillir des informations approfondies sur les résultats de ces systèmes et sur la manière dont ils ont été générés (Lipton, 2018 ; Chazette et al., 2019).

L'intelligence artificielle explicable (en anglais E**X**plainable **A**rtificial **I**ntelligence - **XAI**) représente un ensemble de processus et de méthodes qui visent à produire des systèmes d'IA plus transparents, compréhensibles et explicables, sans diminuer leurs performances et leur précision (Adadi et Berrada, 2018 ; Arrieta et al., 2020 ; Langer et al., 2021). Elle aborde ainsi la responsabilité numérique et les aspects sociaux, éthiques et écologiques de l'utilisation des systèmes d'information (Sovrano et al., 2022). Au moment de la rédaction de ce document (fin 2023), l'IA allait bientôt être réglementée dans l'Union européenne (UE) par le biais de la loi sur l'IA et au Canada par le biais de la loi sur l'intelligence artificielle et les données, qui fait partie du projet de loi C-27 plus large, déposé par le gouvernement du Canada en juin 2022. La réglementation européenne contiendra des exigences en matière d'IA explicable (IAE). Par son soucis de transparence, la IAE viendrait impacter positivement les enjeux éthiques des systèmes d'IA. En effet, plusieurs enjeux de l'IA ont pu être identifiés (cf. Fjeld et al., 2020; Hagendorff, 2020) et nous nous concentrerons ici sur les trois principes éthiques les plus discutées, comme le suggère Hagendorff (2020) : 1) *la protection de la vie privée* ; 2) *l'équité* ; et 3) *la responsabilité*. En mettant l'accent sur ces trois principes et en considérant l'impact potentiellement positif de la IAE sur l'éthique dans les systèmes d'IA, nous répondrons à la question suivante : *La IAE peut-elle mitiger les enjeux éthiques liés à la protection de la vie privée, à l'équité et à la responsabilité dans les systèmes d'IA ?*

Notre analyse suggère que la IAE représente un catalyseur d'une IA plus éthique en identifiant et en localisant les violations de la vie privée et de l'équité ainsi que les parties responsables. Cependant, une action humaine supplémentaire est nécessaire pour traiter et atténuer les enjeux identifiés.

8.2 La prochaine génération de l'IA

Un programme lancé par la Defense Advanced Research Projects (DARPA) des États-Unis en mai 2017 a suscité l'intérêt et la recherche dans le domaine de la IAE (Adamson, 2022 ; Rawal et al., 2022). Le programme intitulé « Broad Agency Announcement : Explainable Artificial Intelligence » avait pour objectif d'encourager la recherche et le développement dans le domaine de la IAE (DARPA, 2016 ; Gunning et Aha, 2019). Cela a déclenché un intérêt mondial (Arrieta et al., 2020) dans un domaine de recherche multidisciplinaire (Langer et al., 2021). La recherche sur la IAE couvre un large éventail de disciplines, allant de l'IA, de l'AA, de la science des données et de l'interaction homme-machine aux sciences sociales et cognitives, à la psychologie et à la philosophie (Adadi et Berrada, 2018 ; Miller, 2019 ; Brunotte et al., 2022a) et cherche à rendre les systèmes IA , ses processus et décisions compréhensibles pour les humains (Gunning et Aha, 2019; Langer et al., 2021).

Aujourd'hui, la pertinence de la IAE est reconnue par de nombreuses parties prenantes, notamment des experts en AA, des régulateurs, des avocats, des philosophes et des futurologues (Mittelstadt et al., 2019). Elle englobe bien plus que quelques méthodes technologiques individuelles et est considérée comme un mouvement et une partie de la prochaine génération de développement de la IAE (Adadi et Berrada, 2018 ; Adamson, 2022). Les techniques de IAE sont développées et employées dans de nombreux secteurs cruciaux, notamment les soins de santé, l'armée, les transports, le droit, la sécurité et la finance (Adadi et Berrada, 2018 ; Arrieta et al., 2020).

8.3 La terminologie et les principes de base de l'IA explicable (IAE)

Afin de bien comprendre les principes de base et la terminologie de l'IA explicable, les définitions des termes et leurs limitations sont cruciales pour la suite de cette analyse. (Gilpin et al., 2018 ; Rawal et al., 2022). Les définitions données ci-dessous s'inscrivent dans le contexte de l'IA.

Le terme *compréhension* désigne « l'action de comprendre le sens, le fonctionnement, la nature, etc. de quelque chose » (Dictionnaire Larousse, s.d.). En termes d'IA, la compréhension fait référence à la connaissance, par exemple, du fonctionnement interne des mécanismes d'un modèle d'IA et de la prédiction des résultats. *La compréhensibilité* décrit la capacité d'un modèle à permettre la compréhension par les humains (Arrieta et al., 2020).

La transparence fait référence à la visibilité des informations (Turilli et Floridi, 2009). Un système d'IA est dit transparent si son fonctionnement interne, ses compositions et ses données d'entraînement sont visibles pour les humains, c'est-à-dire que l'on peut "voir à travers" (Chazette et al., 2019). La transparence permet aux humains de comprendre intrinsèquement les informations visibles (Rawal et al., 2022), qu'il s'agisse de l'ensemble du système ou de certains de ses composants (Lipton, 2018).

L'explicabilité est une autre façon de permettre la compréhension (Chazette et al., 2019 ; Miller, 2019). Elle s'exprime activement en fournissant et en échangeant des raisons et des informations sur le fonctionnement d'un système ou son comportement (Chazette et al., 2019 ; Mittelstadt et al. 2019). Ainsi, une explication agit comme un pont entre les humains et un décideur, leur permettant de communiquer efficacement (Guidotti et al., 2018).

Enfin, le concept d'*interprétabilité* fait référence à la mesure dans laquelle un agent est capable d'acquérir et d'utiliser les informations véhiculées par les explications fournies par le système, ainsi que les informations rendues disponibles par le niveau de transparence du système (Tomsett et al., 2018). Par conséquent, l'information fournie à l'humain doit être expliquée et transparente (Miller, 2019 ; Arrieta et al., 2020).

8.4 Principales approches de l'IA explicable (IAE)

Selon la DARPA (2016), l'idée derrière la IAE est de développer des techniques avancées d'apprentissage automatique qui génèrent des modèles facilement explicables. Ces modèles, associés à des techniques d'explication efficaces, permettront aux utilisateurs de comprendre, de faire confiance et de gérer plus efficacement la nouvelle génération de systèmes d'IA. Pour atteindre ces objectifs, une grande variété de techniques et d'outils IAE ont été développés au cours des dernières années (Adadi et Berrada, 2018 ; Arrieta et al., 2020). Pour catégoriser ces approches, plusieurs classifications existent, qui ne sont ni mutuellement exclusives ni exhaustives (Adadi et Berrada, 2018). Les spécialistes de l'IA font une distinction entre les systèmes d'IA qui peuvent être conçus de manière transparente, et ceux qui ne le peuvent pas (sans perte de performance) (Arrieta et al., 2020). Cette classification a été introduite par Lipton (2018) et est fréquemment reprise dans la littérature.

D'une part, un système peut être conçu de manière à être transparent en lui-même - on parle alors de *transparence* (Lipton, 2018 ; Rawal et al., 2022). Les systèmes transparents sont

comme des boîtes de verre (Rai, 2020) dont le fonctionnement interne et les composants sont visibles pour les humains (Lipton, 2018 ; Tomsett et al., 2018). Les humains peuvent extraire des informations directement afin d'améliorer leur compréhension (Langer et al., 2021). Toutefois, la complexité et les performances des modèles transparents sont généralement limitées (Adamson, 2022).

D'autre part, un système et ses décisions peuvent être rendus compréhensibles par l'application d'un modèle ou d'un composant IAE externe, moins complexe, qui fournit une sorte d'explication pour l'aspect en question - ce que l'on appelle *l'explicabilité post-hoc* (Lipton, 2018 ; Arrieta et al., 2020 ; Rawal et al., 2022). Cela est particulièrement nécessaire pour les systèmes d'IA/AA très profonds et puissants qui ne peuvent pas être rendus transparents sans perte de performance ou de précision (Adadi et Berrada, 2018 ; Adamson, 2022). Il existe plusieurs méthodes d'explication qui peuvent être appliquées à un éventail de systèmes d'IA complexes (Rawal et al., 2022), telles que les explications textuelles, visuelles et basées sur des exemples (Arrieta et al., 2020 ; Rawal et al., 2022).

8.5 L'éthique dans l'IA

L'émergence et les progrès des systèmes d'IA s'accompagnent d'enjeux éthiques croissants dans de nombreux domaines critiques tels que la santé ou le droit (Zhou et al., 2022) en impactant directement la vie humaine (Floridi et al., 2018 ; Zhou et al., 2022). Le besoin d'ajouter de l'éthique dans les systèmes d'IA se fait donc sentir de plus en plus (Arrieta et al., 2020 ; Zhou, 2019).

L'éthique est une sous-discipline de la philosophie (Schlagwein et Willcocks, 2023) qui se réfère à un système de lignes directrices et de principes moraux sur la manière de se comporter dans des situations critiques et dilemmatiques - clarifiant ce qui est bon pour les individus et le grand public (cf. Anderson et Anderson, 2007 ; Santosh et Wall, 2022). L'éthique en tant que science comporte plusieurs domaines majeurs tels que la *métaéthique* (qui s'intéresse à la nature et à la signification de l'éthique), *l'éthique normative* (qui détermine les normes et les valeurs de l'action morale) et *l'éthique appliquée* (qui applique l'éthique à des situations spécifiques, parfois historiquement nouvelles, ou à des domaines d'actions possibles) (Melchert, 2014).

L'éthique de l'IA est une forme d'éthique appliquée, qui s'intéresse aux enjeux éthiques qui se posent dans le contexte des systèmes d'IA (Hagendorff, 2020). L'éthique doit être intégrée dans les systèmes d'IA et les enjeux éthiques doivent y être abordés (Zhou, 2019 ; Baldi et Oliveira, 2022). Des lignes directrices éthiques doivent être établies pour définir les principales exigences et principes éthiques importants pour garantir un comportement éthique dans les systèmes d'IA (cf. Jia et Zhang, 2022). Il existe un certain nombre d'approches techniques et non techniques pour permettre des systèmes d'IA éthiques. Certaines approches techniques sont, par exemple, les méthodes d'explication comme la IAE ou l'éthique par la conception (AI HLEG, 2019). Selon Jia et Zhang (2022), une approche non technique largement adoptée consiste à définir des lignes directrices éthiques claires. Dans son examen systématique de la littérature dans le domaine de l'éthique de l'IA, Hagendorff (2020) analyse et compare vingt-deux lignes directrices éthiques différentes et leurs principes. Son analyse a montré que trois principes, la protection de la vie privée, l'équité/non-discrimination/justice et la responsabilité, apparaissent dans 80 % de la littérature analysée (Hagendorff, 2020). D'autres méta-études aboutissent à des conclusions similaires (Fjeld et al., 2020 ; Khan et al., 2022). Compte tenu de leurs effets négatifs potentiels sur l'humanité, la nécessité de systèmes d'IA éthiques est plus importante que jamais (Guidotti et al., 2018).

8.5.1 La vie privée

En général, le concept de vie privée signifie avoir le contrôle de ses propres données personnelles, de leur accès et de leur utilisation (Brunotte et al., 2023 ; Santosh et Wall, 2022). La définition de la vie privée dans le contexte des systèmes d'IA est cruciale, car elle reconnaît le droit d'une personne à contrôler ses données. En tant que propriétaire des données, l'individu peut fixer des limites et décider des aspects de la vie privée qu'il souhaite partager avec les collecteurs de données et les utilisateurs. Il s'agit notamment de déterminer qui peut accéder aux données et à quel moment, et de veiller au respect de la vie privée (Brunotte et al., 2023).

La protection de la vie privée est progressivement entravée, par exemple, par les applications pour les maisons intelligentes (la configuration d’une maison dans laquelle les appareils et dispositifs peuvent être automatiquement contrôlés à distance depuis n'importe quel endroit

disposant d'une connexion internet, à l'aide d'un appareil mobile ou d'un autre appareil en réseau) ou par les technologies liées à la santé (Brunotte et al., 2022b), car les données à caractère personnel font partie des opérations de fusion, de traitement et d'interprétation (IEEE, 2017). Souvent, les personnes concernées ne sont même pas conscientes des risques qu'elles encourent en matière de protection de la vie privée, ni du type de données collectées, par qui, dans quel but et où elles sont stockées. Les applications de géolocalisation collectent généralement des informations sur les lieux fréquemment visités, tels que le domicile ou le lieu de travail. L'alimentation en données des systèmes d'intelligence artificielle présente de réelles menaces pour la vie privée et, par conséquent, la vie privée est une question d'éthique sérieuse (Rawal et al., 2022).

Ainsi, pour garantir le principe éthique de la protection de la vie privée, une série d'exigences doivent être remplies par les systèmes informatiques et les systèmes d'IA. Les personnes concernées doivent donner leur autorisation et être pleinement informées de l'utilisation et des pratiques relatives à leurs données, tout en conservant un contrôle total sur celles-ci (Abrassart et al., 2018 ; Brunotte et al., 2022b). Ils doivent pouvoir accéder à leurs données personnelles, les rectifier ou les supprimer à tout moment (Abrassart et al., 2018 ; Fjeld et al., 2020 ; Hagendorff, 2020). L'accès aux données doit être limité aux personnes ayant des raisons d'accès appropriées, tandis que la qualité et l'intégrité des données doivent être garanties, y compris un traitement correct, transparent et équitable des données (Abrassart et al., 2018 ; AI HLEG, 2019).

8.5.2 L'équité

En association avec les systèmes d'IA, l'équité est caractérisée comme l'absence de toute forme de partialité ou de favoritisme envers des individus ou des groupes sur la base d'attributs personnels sensibles et indifférents au contexte (Mehrabi et al., 2022) ou comme une condition d'impartialité envers une personne ou un groupe (Agarwal et al., 2022). Garantir l'équité des décisions des systèmes d'IA est un autre défi crucial de l'éthique de l'IA (Saxena et al., 2019 ; Zhou, 2019).

Pour définir l'équité, on évalue une large série de notions, d'interprétations et de mesures qui peuvent être différentes selon la subjectivité personnelle et le contexte (Franklin et al., 2022). Cependant, en raison de l'absence d'une définition unifiée, de l'ambiguïté entre les

notions et les mesures, et de la quantité de littérature existante, l'équité est difficile à mesurer (Franklin et al., 2022). Les systèmes d'IA peuvent intégrer divers biais (Franklin et al., 2022). Ces biais peuvent apparaître à n'importe quel stade du cycle de vie d'un système d'IA (Agarwal et al., 2022). Les trois sources potentielles de biais sont les données, les individus et le système d'IA lui-même (Stinson, 2022).

Les quantités de données utilisées pour la formation des systèmes d'IA peuvent être biaisées, défectueuses ou non représentatives (Fjeld et al., 2020). Il est important d'en être conscient lors de l'analyse et l'interprétation des données pour s'assurer que les conclusions sont exactes et impartiales (Chakraborty et al., 2021). D'autant plus que les données reproduisent inévitablement les jugements humains et les modèles sociaux (Jia et Zhang, 2022). Par conséquent, les décisions prises par les systèmes d'IA peuvent non seulement recréer ces biais, mais aussi les aggraver (Langer et al., 2021). Par exemple, le fait de ne pas utiliser des ensembles de données totalement représentatifs ou de ne pas prendre en compte des caractéristiques dites « sensibles » ou « protégées », telles que la race ou le sexe, peut donner lieu à des résultats injustes ou discriminatoires (Saxena et al., 2019 ; Zhou 2019).

8.5.3 La responsabilité

Selon les lignes directrices en matière d'IA éthique, à ce stade, seuls les êtres humains peuvent être tenus pour responsables des dommages, et non les systèmes d'IA (Abrassart et al., 2018; Wieringa, 2020). Selon Loi et Spielkamp (2021), le concept de responsabilité lui-même et ses différentes formes et dimensions manquent de clarté. Bovens (2007) définit la responsabilité, ou être responsable, comme l'action, de la part de la partie responsable d'une situation préjudiciable, d'expliquer et de justifier ses actions auprès des personnes concernées ou de leurs représentants. Les parties affectées ont le droit de poser des questions et de porter des jugements, et la partie responsable peut être confrontée à des conséquences. Selon cette définition, la responsabilité est de nature relationnelle et comprend cinq éléments clés : l'acteur, le forum, la relation, le contenu et les critères du compte rendu, et les conséquences (Wieringa, 2020). En appliquant la responsabilité relationnelle au contexte de l'IA, un système d'IA ainsi que d'autres parties prenantes, peuvent représenter les acteurs, tandis que les décisions prises par un système d'IA peuvent être considérées comme le contenu et les

critères du compte rendu ; c'est ce que l'on appelle la « responsabilité algorithmique » (Wieringa, 2020).

Les trois mêmes phases sont pertinentes pour tout compte rendu efficace : *information*, *explication/justification* et *conséquences* (Bovens, 2007). Dans la phase d'information, les informations pertinentes sont mises à la disposition du forum par l'acteur qui l'informe de son comportement. La deuxième phase comprend l'interrogation de l'acteur par le forum et la demande d'explications et de justifications concernant le comportement. Ces explications et justifications doivent être claires et compréhensibles (Busuioc, 2021). Dans cette étape, la responsabilité est étroitement liée au concept de *l'obligation de rendre des comptes*, la responsabilité faisant référence à la fois à *la capacité* et à *la volonté* de l'acteur de fournir au forum les raisons et les explications de ses actions, ainsi qu'au *droit du forum* de demander ces raisons (Bovens, 2007 ; Busuioc, 2021). Enfin, la troisième phase identifi les *conséquences* qui doivent être imposées ou du moins possibles (Busuioc, 2021).

La littérature suggère que les systèmes autonomes comme les véhicules ou les applications de diagnostic médical peuvent prendre des décisions qui doivent être prises en compte par quelqu'un (cf. Santosh et Wall, 2022). Par exemple, en 2018, une voiture autonome Uber a provoqué un accident mortel (Wakabayashi, 2018). Bien qu'il y ait eu un conducteur dans la voiture, celle-ci conduisait de manière autonome au moment de l'accident. Cet exemple souligne à lui seul la nécessité d'établir des mécanismes pour garantir la responsabilité legale (Cooper et al., 2022 ; Henriksen et al., 2021) et la responsabilité algorithmique.

Néanmoins, avec la prévalence des systèmes d'IA, il devient de plus en plus problématique de localiser et d'attribuer les responsabilités (Langer et al., 2021). Nissenbaum (1996) identifie quatre obstacles qui compliquent l'attribution des responsabilités avec l'avènement des systèmes informatisés : 1. Le problème de la *multiplicité des mains* - il s'agit du fait que le nombre de parties impliquées passe de quelques unes à un système complexe de parties ; 2. l'apparition de *bogues informatiques* ; 3. le fait de blâmer les systèmes informatiques pour leurs décisions et de les utiliser comme « *boucs émissaires* » ; et 4. le défi de la « propriété sans responsabilité légale » (en anglais - ownership without liability). Le défi de la « propriété sans responsabilité légale » se réfère au problème des droits de propriété des systèmes et de leurs composants.

Avec les progrès des systèmes d'IA par rapport aux systèmes informatiques normaux, des défis supplémentaires se posent (Henriksen et al., 2021). La capacité des systèmes d'IA opaques à apprendre continuellement à partir des données plutôt que d'avoir un code écrit explicite aggrave l'attribution de la responsabilité (Bovens, 2007 ; Henriksen et al., 2021). Un « déficit de responsabilité » apparaît (Lima et al., 2022). L'une des conséquences directes possibles est la prise de décisions autonomes ayant des conséquences négatives dont personne n'est directement responsable - car trop de personnes peuvent avoir contribué (de manière intraçable) au préjudice (cf. Cooper et al., 2022).

En conclusion, afin de répondre aux trois enjeux éthiques majeurs, l'IA explicable doit donc prendre en considération la protection de la vie privée, l'équité et la responsabilité et aider à diminuer les biais issus des systèmes d'IA.

8.6 Comment l'IA explicable (IAE) aborde les enjeux éthiques

Comme mentionné précédemment, les enjeux éthiques sont étroitement liés et exacerbés par l'opacité des systèmes d'IA et la compréhension limitée qui en résulte. Nous rappelons notre principale question de recherche *La IAE peut-elle mitiger les enjeux éthiques liés à la protection de la vie privée, à l'équité et à la responsabilité dans les systèmes d'IA ?* Pour répondre à cette question, nous analysons comment les deux principales techniques de la IAE, *la transparence* et *l'explicabilité a posteriori*, ont un impact sur les enjeux éthiques. Ces deux approches contribuent généralement de manière différente à l'atténuation des enjeux éthiques. Par conséquent, les implications potentielles de la transparence et de l'explicabilité sur les trois enjeux éthiques sont abordées ultérieurement.

8.6.1 La transparence

La transparence peut sembler à première vue l'approche la plus évidente pour résoudre le problème de l'opacité puisque la transparence est décrite comme le contraire de l'opacité. La création de systèmes transparents peut être un outil précieux pour découvrir des enjeux éthiques lorsque les informations nécessaires à l'examen des enjeux éthiques sont rendues visibles (Turilli et Floridi, 2009). En rendant visibles les informations sur le système d'IA, les humains peuvent mieux comprendre et localiser les menaces potentielles pour la vie

privée, l'équité et/ou la responsabilité. Toutefois, la création de systèmes d'IA transparents s'accompagne de difficultés. Tout d'abord, la conception et le développement transparents des systèmes d'IA ne sont généralement pas réalisables sans perte de performance et/ou de précision, à moins de considérer des modèles simplifiés à l'extrême. Réaliser des systèmes d'IA efficaces et transparents est actuellement un objectif contradictoire. Deuxièmement, même si les systèmes d'IA peuvent être rendus transparents, les informations rendues visibles, comme le code, les données ou les processus algorithmiques, sont souvent dénuées de sens et non compréhensibles pour les néophytes (cf. Doran et al., 2017 ; Langer et al., 2021) et parfois même pour les ingénieurs et les experts (Köhl et al., 2019). Cela est dû, d'une part, au manque d'expertise et de connaissances et, d'autre part, à la complexité des systèmes d'IA (Adadi et Berrada, 2018 ; Busuioc, 2021 ; Santosh et Wall, 2022). Par exemple, les néophytes comme les utilisateurs peuvent être en mesure d'inspecter le code d'un système d'IA transparent mais peuvent ne pas comprendre, à l'étape suivante, ce que le code sous-jacent signifie, ce que le système d'IA fait et comment il prend des décisions. Au contraire, les décideurs peuvent disposer d'une expertise suffisante, mais l'énorme quantité de données et de paramètres introduits dans un système d'IA et créés dans ses propres représentations rend souvent impossible la saisie de la complexité des interactions entre les caractéristiques, même lorsqu'elles sont totalement transparentes (Adadi et Berrada, 2018 ; Busuioc, 2021).
Dans ces deux cas, la transparence seule ne permet pas implicitement la compréhension. Par conséquent, la création de systèmes transparents sans perte de caractéristiques importantes n'est généralement pas réalisable et/ou pas suffisante. Ainsi, ni les violations de la vie privée ni les injustices existantes ne peuvent nécessairement être identifiées ou atténuées, et le code transparent ne permet pas nécessairement de rendre des comptes. La transparence n'est pas (encore) prometteuse pour réaliser des systèmes d'IA plus privés, plus équitables ou plus responsables, ou seulement dans des cas exceptionnels. Mais supposons que la transparence soit théoriquement possible, car il n'est pas exclu que des systèmes d'IA très complexes, mais transparents, soient développés à l'avenir. Alors, la transparence pourrait contribuer différemment à la compréhension et à l'atténuation des enjeux éthiques, selon l'un des trois niveaux de transparence : *la simulabilité*, *la décomposabilité* ou la *transparence algorithmique*.

Pour illustrer les différents effets de la transparence sur l'éthique dans les systèmes d'IA, nous prenons l'exemple d'un système de recrutement par IA formé à partir de données historiquement biaisées concernant le sexe. Dans cet exemple, la caractéristique du sexe prise en compte dans le processus décisionnel du système d'IA est sensible et pertinente d'un point de vue éthique. Par conséquent, les décisions du système sont injustes d'un point de vue éthique.

Du point de vue de la *simulabilité, un haut* niveau de transparence global permettrait aux humains de simuler l'ensemble du système de recrutement, y compris tous les composants, paramètres et données d'entrée en une seule fois (cf. Lipton, 2018). Sur la base des données d'un candidat donné en entrée et des paramètres visibles, les humains seraient en mesure de simuler la décision du système de recrutement concernant le candidat. On pourrait alors identifier que la caractéristique du sexe a une influence majeure sur le processus de décision (à supposer que l'expertise nécessaire soit présente). Cette information pourrait être utilisée pour corriger et éliminer la caractéristique du sexe du processus de décision. Par conséquent, la simulabilité pourrait contribuer à atténuer la question éthique de l'équité en la rendant perceptible.

La décomposabilité permettrait de rendre visible séparément chacun des composants du système de recrutement (cf. Lipton, 2018). Par exemple, les données d'entrée, les calculs et les paramètres seraient individuellement transparents. Ce niveau de transparence ne mènerait pas directement à l'identification de l'enjeu d'équité (biais de la donnée sexe) car les paramètres transparents sont considérés séparément, et non dans un contexte logique avec le processus de calcul et les données d'entrée. Dans le cas où des composants individuels du modèle peuvent être combinés par des humains pour comprendre la logique cohérente globale, par exemple en cartographiant le paramètre de sexe à la décision algorithmique, l'injustice sous-jacente pourrait être identifiée. Cela nécessiterait cependant une interprétation humaine plus approfondie des composants visibles. Par conséquent, l'effet de la décomposabilité sur les systèmes d'IA éthiques dépend de la possibilité de combiner les informations des composants individuellement transparents pour comprendre la logique sous-jacente globale et mitiger l'enjeu éthique.

L'activation de la *transparence algorithmique seule* permettrait de divulguer le processus algorithmique qui sous-tend la décision de recrutement. Le fait de pouvoir inspecter le

processus algorithmique seul ne rend pas visibles les paramètres ou les données d'entrée. Par conséquent, le seul fait d'assurer la transparence du processus ne permet pas de découvrir que la caractéristique du sexe est fortement pondérée dans la décision de recrutement. La transparence algorithmique ne suffirait pas à identifier la pertinence de cette caractéristique et pourrait ne pas conduire à des systèmes d'IA plus éthiques puisqu'elle ne rend visible que l'algorithme et ne contient pas d'informations sur les données d'entrée ou les paramètres.

Par conséquent, si elle est réalisable et comprise, *la transparence* totale pourrait être une condition préalable importante pour atténuer les problèmes éthiques. Plus les parties des systèmes d'IA sont rendues transparentes, plus il y a d'informations visibles qui peuvent être inspectées ; les violations de la vie privée et de l'équité ainsi que les parties responsables peuvent être identifiées. Les problèmes éthiques deviennent apparents avec la transparence. Néanmoins, ils doivent être résolus par les humains.

Étant donné qu'il est peu probable que la transparence soit disponible rapidement ou qu'elle permette la compréhension, la technique IAE de transparence n'est que dans des cas exceptionnels une solution réalisable pour obtenir des systèmes d'IA plus privés, plus équitables et plus responsables. Cela équivaut à la nécessité d'expliquer (cf. Doran et al., 2017 ; Langer et al., 2021 ; Santosh et Wall, 2022). Les explications peuvent être nécessaires soit comme substitut dans le cas où la transparence n'est pas réalisable, soit en complément de la transparence dans le cas où la transparence ne permet pas implicitement la compréhension. C'est pourquoi la plupart des modèles IAE actuels sont basés sur des techniques d'explicabilité a posteriori (Adamson, 2022).

8.6.2 L'explicabilité a posteriori

L'explicabilité a posteriori est considérée comme une technique pratique pour favoriser la compréhension et satisfaire aux exigences éthiques d'un système (Chazette et al., 2021). Mais comment ? En ce qui concerne la protection de la vie privée, l'équité et la responsabilité, les explications peuvent révéler des informations explicatives utiles d'un système d'IA (autrement) opaque concernant, par exemple, son fonctionnement, ses processus ou ses entrées et sorties. Les humains peuvent utiliser toutes ces informations pour comprendre les systèmes d'IA, un peu comme un mode d'emploi. Les explications et la compréhension qui en résultent sont utiles pour contrôler la conformité des systèmes d'IA avec les trois grands

principes éthiques. Les humains peuvent contrôler ou justifier directement les décisions des systèmes d'IA ou peuvent indirectement tirer des conclusions causales ou identifier des modèles à partir des explications (Langer et al., 2021) pour obtenir d'autres informations et détails sur le système d'IA et ses décisions.

Pour déterminer comment et dans quelle mesure les explications contribuent à atténuer les problèmes éthiques liés à la protection de la vie privée, à l'équité et à la responsabilité, il est essentiel d'examiner la compréhension réelle obtenue (Langer et al., 2021). Un aspect doit non seulement être expliqué, mais aussi compris pour contribuer éventuellement à des systèmes d'IA plus privés, équitables et responsables. Par exemple, si une explication peut révéler des informations sur une caractéristique sensible prise en compte dans le processus décisionnel du système d'IA, mais que l'inspecteur humain ne comprend pas les informations explicatives, l'injustice de l'utilisation de la caractéristique sensible ne peut pas être identifiée et l'explication elle-même ne contribue pas à des systèmes d'IA plus équitables.

Les explications a posteriori peuvent révéler des informations nécessaires pour reconnaître l'injustice, les atteintes à la vie privée et les personnes responsables du préjudice causé. Toutefois, l'atténuation des enjeux liés à la protection de la vie privée ou à l'équité doivent toujours être pris en charge par les humains. Par conséquent, fournir une explication peut aider à justifier et à contrôler un comportement éthique ou à identifier des problèmes éthiques concernant la vie privée, l'équité et la responsabilité, mais ne peut pas directement atténuer ces problèmes. Néanmoins, l'utilisation d'explications permet de prévenir ou d'atténuer les comportements contraires à l'éthique.

En général, plus les informations sont accompagnées d’explications adéquates, plus les humains peuvent comprendre, contrôler et améliorer les systèmes d'IA, et plus les questions et enjeux éthiques peuvent être identifiés et diminués par les humains. Cependant, dans la pratique de la IAE, les explications sont le plus souvent données avec des termes très techniques qui ne peuvent être compris qu'avec un certain niveau d'expertise (Floridi et al., 2018). La manière de rendre ces explications compréhensibles pour tous reste à déterminer à l'avenir (Cortese et al., 2022).

8.7 Conclusion

Les systèmes d'IA, qui deviennent de plus en plus omniprésents dans notre société, redéfinissent nos valeurs, nos propriétés et nos interactions quotidiennes. Cependant, cette intrusion apporte son lot d'enjeux éthiques. Dans notre analyse, nous avons introduit l'approche de l'intelligence artificielle explicable (IAE) comme pouvant être un catalyseur d'une IA plus éthique. En effet, nous avons déterminé que la IAE peut contribuer à mitiger les enjeux éthiques liés à la vie privée, à l'équité et à la responsabilité, mais elle ne peut pas les modérer à elle seule. Pour tirer pleinement bénéfice de la IAE, les informations divulguées doivent être adaptées et comprises. Il faut choisir le bon type et le bon niveau de technique de IAE et divulguer le bon aspect au(x) bon(s) destinataire(s) pour contribuer à la protection de la vie privée, à l'équité et à la responsabilité. Le tableau 1 résume les avantages et les inconvénients de la IAE pour la protection de la vie privée, l'équité et la responsabilité.

La première contribution que la IAE peut apporter, se trouve dans le domaine de la transparence. En utilisant l'approche de transparence et ses trois niveaux (simulabilité, décomposabilité et transparence algorithmique), la IAE permet aux systèmes d'IA d'être plus compréhensibles aux humains et d'identifier les menaces qu'ils représentent sur les enjeux éthiques. Malheureusement la création de systèmes transparents sans perte de caractéristiques importantes n'est généralement pas réalisable et/ou pas suffisante.

Donc, pour aller plus loin, la deuxième contribution que la IAE aborde, est celle de l'explicabilité à posteriori. Avec cette approche, les systèmes d'IA peuvent être conrôler, les décisions justifier et il est possible aux humains de tirer des conclusions causales et identifier des explications des comportements. En expliquant et en rendant plus visible, lisible et compréhensible l'information, la IAE permet d'atténuer de manière indépendante les causes des problèmes éthiques.

Malgré tout, le dernier enjeu, la responsabilité, pourrait être mieux prise en compte par la IAE si l'ensemble du processus de création des systèmes d'IA, y compris leur conception, leur développement et leur mise en œuvre, était documenté. La IAE pourrait utiliser ces informations pour localiser plus précisément les parties responsables tout au long du processus de création des systèmes d'IA.

Tableau 8.1 Effets de la IAE sur la vie privée, l'équité et la responsabilité

Avantages IAE	Principes éthiques	Inconvénients de la IAE
+ Identification des violations de la vie privée et localisation de leurs causes + Justification de la protection de la vie privée des systèmes d'IA + + Sensibilisation au respect de la vie privée	Vie privée	– Exposition des données personnelles – Risque de classification incorrecte des systèmes d'IA comme étant privés
+ Identification de l'iniquité et localisation de ses causes + Justification de l'équité des systèmes d'IA	Équité	– Risque de classification incorrecte des systèmes d'IA comme étant équitables
+ Aide à l'identification des parties responsables + Fournit des informations significatives pour la reddition de comptes	Responsabilité	– La transparence et la documentation du processus de création des systèmes d'IA ne sont pas suffisantes.

En conclusion, la IAE n'en est qu'à ses premiers balbutiements et il reste une grande place à l'amélioration. Comme il n'existe pas encore d'accord commun sur la définition d'une « explication » (Lipton, 2018; Köhl et al., 2019), ses caractéristiques requises (Guidotti et al., 2018) ou son déploiement dans la pratique (Hafermalz et Huysman, 2021), une grande confusion règne entre une IA « explicable » et une IA réellement « explicative » (Sovrano et al., 2021). La IAE sera-t-elle en mesure d'unifier les recherches entre les sciences de l'explication, qui comprennent les sciences cognitives et sociales, la philosophie et le droit ? En attendant, la IAE représente une belle approche pour apporter plus de transparence dans des systèmes complexes et opaques.

8.8 Bibliographie

Abrassart, C., Bengio, Y., Chicoisne, G., de Marcellis-Warin, N., Dilhac, M.-A., Gambs, S., … et Voarino, N. (2018). « La Déclaration de Montréal pour un développement responsable de l'intelligence artificielle », Université de Montréal. (https://declarationmontreal-iaresponsable.com/la-declaration/, consulté le 16 octobre 2023).

Adadi, A., et Berrada, M. (2018). « Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) », IEEE Access (6), p. 52138–52160.

Adamson, G. (2022). Ethics and the explainable artificial intelligence (XAI) movement. TechRxiv, 1-13.

Agarwal, A., Agarwal, H., et Agarwal, N. (2022). « Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems », AI and Ethics (3), 267–279.

AI HLEG. (2019). « Ethics guidelines for trustworthy AI », Report for the European Commission by the High-Level Expert Group on Artificial Intelligence (AI HLEG). Report no. B-1049. Brussels, Belgium. (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, consulté le 16 octobre 2023).

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... et Herrera, F. (2023). « Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence ». Information Fusion, 99, 101805.

Anderson, S. L., et Anderson, M. 2007. « Machine Ethics: Creating an Ethical Intelligent Agent », AI Magazine (28), p. 15-26.

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … et Herrera, F. (2020). « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », Information Fusion (58), p. 82-115.

Baldi, V., et Oliveira, L. (2022). « Challenges to incorporate accountability into artificial intelligence », Procedia Computer Science (204), p. 519-523.

Bovens, M. (2007). « Analysing and Assessing Accountability: A Conceptual Framework », European Law Journal (13:4), p. 447-468.

Brunotte, W., Chazette, L., Klös, V., et Speith, T. (2022a). « Quo Vadis, Explainability? – A Research Roadmap for Explainability Engineering », in Requirements Engineering: Foundation for Software Quality, V. Gervasi and A. Vogelsang (eds.), Springer International Publishing, p. 26–32.

Brunotte, W., Chazette, L., Kohler, L., Klunder, J., et Schneider, K. (2022b). « What About My Privacy? Helping Users Understand Online Privacy Policies », Actes de conférence de l'International Conference on Software and System Processes and International Conference on Global Software Engineering, Pittsburgh PA USA, p. 56-65.

Brunotte, W., Specht, A., Chazette, L., et Schneider, K. (2023). « Privacy Explanations - A Means to End-User Trust », Journal of Systems and Software (195), 111545.

Busuioc, M. (2021). « Accountable Artificial Intelligence: Holding Algorithms to Account », Public Administration Review (81:5), p. 825-836.

Chakraborty, J., Majumder, S., et Menzies, T. (2021). « Bias in machine learning software: Why? How? What to do? », Actes de conférence de 29th ACM Joint Meeting on

European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, p. 429-440.
Chazette, L., Karras, O., et Schneider, K. (2019). « Do End-Users Want Explanations? Analyzing the Role of Explainability as an Emerging Aspect of Non-Functional Requirements », Actes de conférence de l'IEEE 27th International Requirements Engineering Conference, Jeju Island, Korea, p. 223-233.
Chazette, L., Brunotte, W., et Speith, T. (2021). « Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue », Actes de conférence de l'IEEE 29th International Requirements Engineering Conference, Notre Dame, IN, USA, p. 197-208.
Cooper, A. F., Moss, E., Laufer, B., et Nissenbaum, H. (2022). « Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning », Actes de conférence de l'ACM Conference on Fairness, Accountability, and Transparency, Seoul, Korea, p. 864-876.
Cortese, J. F. N. B., Cozman, F. G., Lucca-Silveira, M. P., et Bechara, A. F. (2022). « Should explainability be a fifth ethical principle in AI ethics? », AI and Ethics (3), p. 123-134.
DARPA (Defense Advanced Research Projects Agency). (2016). Broad Agency Announcement Explainable Artificial Intelligence (XAI), Arlington, VA. (https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf. consulté le 10 octobre 2023).
Doran, D., Schulz, S., et Besold, T. R. (2017). « What Does Explainable AI Really Mean? A New Conceptualization of Perspectives », arXiv preprint arXiv:1710.00794.
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., et Srikumar, M. (2020). « Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI », Berkman Klein Cemter Research Publication, (2020-1) Center for Internet & Society, (http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420, consulté le 14 octobre 2023).
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, … et Vayena, E. (2018). « AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations », Minds and Machines (28:4), p. 689-707.
Franklin, J. S., Bhanot, K., Ghalwash, M., Bennett, K. P., McCusker, J., and McGuinness, D. L. 2022. « An Ontology for Fairness Metrics », Actes de conférence de l'AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, p. 265-275.
Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., et Kagal, L. (2018). « Explaining Explanations: An Overview of Interpretability of Machine Learning », Actes de conférence de l'IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, p. 80–89.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). « A Survey of Methods for Explaining Black Box Models », ACM Computing Surveys (51:5), p. 1-42.
Gunning, D., et Aha, D. 2019. « DARPA's Explainable Artificial Intelligence (XAI) Program », AI Magazine (40:2), p. 44-58.
Hafermalz, E., et Huysman, M. 2021. « Please Explain: Key Questions for Explainable AI research from an Organizational perspective », Morals & Machines (1:2), p. 10–23.
Hagendorff, T. 2020. « The Ethics of AI Ethics: An Evaluation of Guidelines », Minds and Machines (30:1), p. 99-120.
Henriksen, A., Enni, S., et Bechmann, A. (2021). « Situated Accountability: Ethical Principles, Certification Standards, and Explanation Methods in Applied AI », Actes de

conférence de l'AAAI/ACM Conference on AI, Ethics, and Society, Online Event, Association for Computing Machinery, p. 574-585.

IEEE (The Institute of Electrical and Electronics Engineers) (2017). « Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems », The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Version 2. (https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf, consulté le 10 octobre 2023).

Jia, K., et Zhang, N. (2022). « Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines », Electronic Markets (32:1), p. 59–71.

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., et Azeem Akbar, M. (2022). « Ethics of AI: A Systematic Literature Review of Principles and Challenges », Actes de conférence de l'International Conference on Evaluation and Assessment in Software Engineering 2022, Gothenburg, Sweden, p. 383–392.

Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., et Bohlender, D. (2019). « Explainability as a Non-Functional Requirement », Actes de conférence de l'IEEE 27th International Requirements Engineering Conference, Jeju Island, SKorea, p. 363–368.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., et Baum, K. (2021). « What Do We Want From Explainable Artificial Intelligence (XAI)? -- A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research », Artificial Intelligence (296), 103473.

Lima, G., Grgić-Hlača, N., Jeong, J. K., et Cha, M. (2022). « The Conflict Between Explainable and Accountable Decision-Making Algorithms », Actes de conférence de l'ACM Conference on Fairness, Accountability, and Transparency, Seoul, Korea, p. 2103–2113.

Lipton, Z. C. (2018). « The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery », Queue (16:3), p. 31–57.

Loi, M., et Spielkamp, M. (2021). « Towards Accountability in the Use of Artificial Intelligence for Public Administrations », Actes de conférence de l'AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, p. 757–766.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., et Galstyan, A. (2022). « A Survey on Bias and Fairness in Machine Learning », ACM Computing Surveys (54:6), p. 1-35.

Melchert, N. (2014). The Great Conversation: A Historical Introduction to Philosophy [7ed.]. Oxford University Press.

Miller, T. (2019). « Explanation in artificial intelligence: Insights from the social sciences », Artificial Intelligence (267), p. 1-38.

Mittelstadt, B., Russell, C., et Wachter, S. (2019). « Explaining Explanations in AI », Actes de conférence de la Conference on Fairness, Accountability, and Transparency, Atlanta, USA, p. 279–288.

Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., ... et Yelis, M. (2022). « Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges », Mathematics, (10:15), 2552.

Nissenbaum, H. (1996). « Accountability in a computerized society », Science and Engineering Ethics (2:1), p. 25-42.

Rai, A. (2020). « Explainable AI: From Black Box to Glass Box », Journal of the Academy of Marketing Science (48:1), p. 137-141.

Rawal, A., McCoy, J., Rawat, D. B., Sadler, B. M., et Amant, R. St. (2022). « Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives », IEEE Transactions on Artificial Intelligence (3:6), p. 852-866.

Santosh, K., et Wall, C. (2022). AI, Ethical Issues and Explainability - Applied Biometrics, Springer Nature Singapore, p. 1-46.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., et Liu, Y. (2019). « How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic

Schlagwein, D., et Willcocks, L. (2023). « 'ChatGPT et al.': The Ethics of Using (Generative) Artificial Intelligence in Research and Science », Journal of Information Technology, (38:3), p. 232-238.

Sovrano, F., Vitali, F., et Palmirani, M. (2021). « Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR », dans AI Approaches to the Complexity of Legal Systems XI-XII, V. Rodríguez-Doncel, M. Palmirani, M. Araszkiewicz, P. Casanovas, U. Pagallo and G. Sartor (eds.), Springer International Publishing, p. 169-182.

Stinson, C. (2022). « Algorithms are not neutral: Bias in collaborative filtering », AI and Ethics (2:4), p. 763-770.

Tomsett, R., Braines, D., Harborne, D., Preece, A., et Chakraborty, S. (2018). « Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems », dans ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden.

Turilli, M., et Floridi, L. (2009). « The ethics of information transparency », Ethics and Information Technology (11:2), p. 105-112.

Wakabayashi, D. (2018). « Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam », The New York Times. (https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html, consulté le 16 octobre 2023).

Wieringa, M. (2020). « What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability », Actes de conférence de la Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, p. 1-18.

Zhou, A. (2019). « The intersection of ethics and AI », AI Matters (5:3), p. 64-69.

Zhou, J., Chen, F., et Holzinger, A. (2022). « Towards Explainability for AI Fairness », dans xxAI - Beyond Explainable AI, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller and W. Samek (eds.), Springer International Publishing, p. 375-386.