Dialogue (2025), **64**, 81–99
doi:10.1017/S0012217325000095

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Numéro spécial sur la normativité. À la recherche des normes perdues

Le développement d'une IA explicable : entre principes éthiques généraux et mesures concrètes

Camélia Raymond[1]*, Marc-Kevin Daoust[2] et Sylvie Ratté[3]

[1]Département de génie logiciel et des technologies de l'information, École de technologie supérieure, Montréal, QC, Canada, [2]Département des enseignements généraux, École de technologie supérieure, Montréal, QC, Canada et [3]Département de génie logiciel et des technologies de l'information, École de technologie supérieure, Montréal, QC, Canada
*Autrice-ressource. Courriel : camelia.raymond@ens.etsmtl.ca

Résumé
Les ingénieurs en IA ont besoin de directives applicables pour l'implémentation de principes éthiques dans leurs solutions technologiques. Mais comment y arriver ? Dans cet article, nous prenons le cas du développement de l'intelligence artificielle explicable (XIA) comme point de départ. Sur le plan des mesures concrètes devant être intégrées à l'IA pour la rendre explicable, nous remettons en question l'approche universaliste. Nous proposons une méthodologie normative pour évaluer les mesures de la XIA adaptées à des contextes spécifiques. Cette approche intègre l'éthique dans le développement de l'IA, offrant ainsi une méthode pragmatique pour les ingénieurs, régulateurs et chercheurs en éthique.

Abstract
AI engineers need applicable guidelines for implementing ethical principles into their technological solutions. But how can this be achieved? In this article, we take the development of Explainable Artificial Intelligence (XAI) as our starting point. First, we challenge the universalist approach, the view according to which some measures are necessary or sufficient for XAI in every context. Then, we propose a normative methodology for evaluating XAI measures that is adapted to specific contexts. This approach better integrates ethics into AI development by offering an adaptable and pragmatic method for engineers, regulators, ethical researchers, and decision-makers.

Mots-clés : intelligence artificielle explicable (XIA) ; éthique de l'IA ; adaptabilité contextuelle ; mesures concrètes

Introduction

Les attentes sociales et commerciales envers les ingénieurs logiciels ne se limitent pas à la compétence technique. Dans le monde de l'intelligence artificielle (IA), les ingénieurs doivent aussi incorporer la dimension éthique dans leurs travaux. Historiquement focalisés sur l'innovation technique et la résolution de problèmes complexes, les ingénieurs en IA sont aujourd'hui confrontés à une responsabilité grandissante vis-à-vis des effets sociaux des technologies. Cette évolution est largement motivée par une pression sociale croissante pour un développement technologique responsable, qui considère les impacts sociétaux (Garrett et al., 2020).

Parallèlement à cette prise de conscience, un écart notable persiste entre les compétences techniques acquises par les ingénieurs et leurs connaissances en éthique. Bien que les cursus en ingénierie intègrent des principes éthiques, les exemples concrets liés au développement logiciel, et plus spécifiquement à l'IA, demeurent restreints (Garrett et al., 2020). Les développements récents dans les programmes universitaires tentent de combler cette lacune, mais une grande partie des ingénieurs actuellement en poste n'ont pas bénéficié de cette formation élargie. Cette situation soulève un problème pour les ingénieurs en IA : être tenu responsable des développements technologiques dans le domaine de l'IA, sans en comprendre complètement les enjeux éthiques. Face à ces manques en matière de formation, les ingénieurs en IA s'efforcent de trouver des solutions pratiques pour relever les défis éthiques inhérents à leur domaine.

La nécessité d'augmenter la transparence des IA, c'est-à-dire d'en faciliter la compréhension et l'accès à l'information, s'inscrit dans ce tournant. Un effort collectif émerge pour développer et intégrer des techniques d'intelligence artificielle explicable (XIA). Cet engagement se reflète dans les réglementations proposées par le Canada, les États-Unis ou encore la Commission européenne, qui établissent des exigences spécifiques concernant la XIA. Par exemple, voici comment la Commission européenne définit cette dernière :

> L'explicabilité concerne la capacité d'expliquer à la fois les processus techniques d'un système d'intelligence artificielle et les décisions humaines qui s'y rapportent (par exemple, domaines d'application d'un système d'intelligence artificielle). L'explicabilité technique suppose que les décisions prises par un système d'intelligence artificielle peuvent être comprises et retracées par des êtres humains. [. . .] Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur). (Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies, 2019)

Il y a deux grandes tendances dans la recherche sur la XIA. La première peut être qualifiée d'*universelle*, au sens où elle cherche à établir des principes et des mesures concrètes qui sont valables dans tous les contextes possibles. La seconde peut être qualifiée de *contextuelle*, au sens où elle part du principe que les mesures concrètes à implémenter dans une organisation vont, dans une large mesure, dépendre des

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

caractéristiques ou faits de la situation (par exemple : les types d'utilisateurs de la plateforme et l'environnement dans lequel la solution est déployée).

Cet article poursuit deux objectifs. D'abord, en ce qui a trait à la XIA, nous remettons en question l'approche des solutions universelles. Le rejet des approches universalistes soulève toutefois un problème. Si nous souhaitons bien guider les ingénieurs voulant intégrer l'éthique à leur travail, que pouvons-nous leur dire de plus que « ce qu'il faut faire dépend du contexte » ? En d'autres termes, quels outils pouvons-nous fournir aux ingénieurs pour les aider à naviguer à travers les situations auxquelles ils sont confrontés ? Ce questionnement nous amène ensuite à proposer une méthodologie pratique de réflexion permettant d'évaluer l'efficacité d'une approche de la XIA pour soutenir les ingénieurs dans la prise de décisions éthiques.

Le plan de l'article va comme suit. Dans les sections 1 et 2, nous clarifions ce que nous voulons dire par approches « universaliste » et « contextualiste », et nous présentons la méthodologie de Clinton Castro et al. (2023) permettant de déterminer quelle approche est appropriée à certains principes éthiques. Puis, dans la section 3, nous évaluons trois mesures de la XIA avec cette méthodologie, en nous posant deux questions essentielles pour chaque paire principe éthique/mesure concrète : la mesure est-elle (1) suffisante et (2) nécessaire pour satisfaire au principe ? Nos observations indiquent que, bien qu'une mesure spécifique puisse être utile dans certains contextes, aucune mesure ne semble universellement nécessaire ou suffisante. Cela nous amène à proposer aux ingénieurs une approche devant composer avec l'incertitude générée par les différents contextes. La section 4 expose cette méthodologie, que l'on nomme la matrice normative pour l'évaluation des mesures en XIA. La section 5 présente l'application de cette matrice au moyen de scénarios fictifs dans le secteur de l'aéronautique.

En proposant cette approche, l'article enrichit la littérature sur la XIA en privilégiant l'analyse de mesures d'explicabilité adaptées aux besoins spécifiques des projets plutôt que la recherche d'une solution universelle. En introduisant la matrice normative pour l'évaluation des mesures en XIA, ce travail souligne l'importance d'une analyse contextuelle et offre un cadre méthodologique pour intégrer les considérations éthiques dans le développement de l'IA. Cette démarche vise à fournir aux ingénieurs en IA des outils pour une réflexion pratique sur l'application de l'explicabilité, contribuant ainsi à une meilleure compréhension et mise en oeuvre des principes éthiques dans l'industrie de l'IA.

1. Aperçu des tendances universelles et contextuelles de la XIA

La recherche sur la XIA se divise en deux grandes tendances : les approches universelles et les approches contextuelles. Les approches universelles cherchent à développer des standards et mesures applicables à tous les contextes, en mettant l'accent, par exemple, sur la création d'un lexique commun, de méthodes uniformes, et de critères d'évaluation fixes pour les décisions des modèles d'IA. Par comparaison, les approches contextuelles adaptent l'explicabilité aux spécificités des différents domaines d'application, reconnaissant que les besoins et les définitions de l'explicabilité varient selon le contexte ciblé.

La recherche d'approches universelles constitue un axe majeur des efforts scientifiques actuels. Ces recherches visent, par exemple, à établir un lexique

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

standardisé pour le domaine de la XIA, à développer des méthodes et techniques applicables de manière globale et à concevoir des critères d'évaluation uniformes pour les approches de la XIA. Selon Othman Benchekroun et al. (2020), « La XIA est primordiale pour une IA de qualité industrielle ; cependant, les méthodes existantes ne répondent pas à cette nécessité (industrielle), en partie en raison d'un manque de standardisation des méthodes d'explicabilité » (Benchekroun et al., 2020, p. 1 ; nous traduisons). Les « standards » mentionnés ici font référence à des outils qui devraient être uniformément intégrés dans les pratiques des entreprises pour renforcer la XIA. Pradeep Reddy et Pavan Kumar (2023) identifient le manque de solutions XIA universelles et standardisées comme étant l'un des défis majeurs de l'IA. Leander Weber et al. (2023) ainsi que Christopher J. Anders et al. (2021) proposent une solution universelle de la XIA, alors que Phuong Quynh Le et al. (2023) ainsi que Mohamed Karim Belaid et al. (2022) exposent une méthode qui permet d'évaluer collectivement les solutions de la XIA. Le dénominateur commun de ces programmes de recherche est que les solutions proposées sont censées s'appliquer à tous les contextes imaginables.

La deuxième tendance de recherche, l'approche contextuelle, se présente comme un axe émergent de la XIA. De plus en plus de chercheurs se penchent sur cet axe, arguant que, malgré l'abondance de recherches, l'approche universelle peine à fournir des solutions efficaces. À l'inverse de l'approche universelle, les approches contextuelles cherchent à établir des lexiques spécifiques à des secteurs industriels (médecine, système bancaire, aviation) ou encore des lexiques dont la terminologie permet de définir le contexte. Par exemple, Rune Nyrup et Diana Robinson (2022) proposent une vision contextuelle de la XIA appliquée au domaine médical. Pour eux, les standards d'explicabilité varient en fonction, notamment, du public et des objectifs du système. Meike Nauta et al. (2023) remettent même en question l'utilité de la définition de la XIA, estimant que celle-ci devrait varier en fonction du contexte. Dans la littérature, une critique récurrente formulée contre les algorithmes de la XIA stipule que, puisque ceux-ci sont souvent développés sans une définition claire de leur utilisation spécifique, ils sont inappropriés dans certains contextes. Q. Vera Liao et Kush R. Varshney (2021), Niels van Berkel et al. (2022) ainsi que Heike Felzmann et al. (2020) mettent en évidence que le contexte d'utilisation de ces algorithmes affecte le succès de leur déploiement auprès des utilisateurs. Aussi, Jianlong Zhou et al. (2021) soutiennent qu'il n'existe pas encore d'indicateur reconnu pour la qualité des méthodes d'explication dans la XIA, puisque celle-ci est un concept subjectif et que la qualité perçue d'une explication dépend de l'environnement et des parties prenantes.

La dualité entre les approches universelles et contextuelles de la XIA met en lumière la complexité de développer des explications d'IA qui soient à la fois précises et largement applicables.

2. Approche normative de justification d'une mesure concrète pour un principe éthique appliqué à l'IA

2.1. Des grands principes aux mesures concrètes

La distinction entre les approches universaliste et contextualiste concerne le choix des *mesures concrètes* à mettre en application dans une organisation ou un système. Pour

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

bien comprendre ce point, il faut d'abord faire quelques rappels concernant les démarches d'analyse courantes en éthique des technologies.

Dans la recherche appliquée en éthique des technologies, les avis comprennent généralement au moins trois éléments : (i) une description des faits de la situation, (ii) une description des valeurs en jeu et (iii) des mesures concrètes à implémenter dans une organisation ou un système. Prenons, par exemple, les 19 avis publiés par la Commission de l'éthique en science et en technologie (CEST) de 2003 à 2024. Tous ces avis ont sensiblement la même structure. On y décrit d'abord le fonctionnement d'une technologie, ainsi que les actions déjà posées au Québec et au Canada pour encadrer la technologie en question. Puis, on esquisse différentes valeurs pertinentes dans la situation. Finalement, on propose des mesures concrètes aux décideurs pour bien encadrer la technologie à l'étude (voir, par exemple, CEST, 2021 ; 2023a ; 2023b ; 2023c ; 2024).

Du point de vue de la pratique des ingénieurs, s'en tenir à une description de grands principes devant régir les technologies n'est pas éclairant. Il faut relier ces principes à des mesures concrètes. Les valeurs devant guider le développement de l'IA sont bien documentées dans la littérature. Plusieurs contributions récentes affirment que l'IA devrait encourager l'exercice de notre autonomie et de notre agentivité (Rubel et al., 2021), favoriser notre compréhension de ses décisions (Fleisher, 2022), tendre à développer les vertus propres à nos fonctions (van Wynsberghe, 2016), respecter notre vie privée (Nissenbaum, 2009), ne pas accentuer les dynamiques d'oppression présentes dans la société (Noble, 2018), et ainsi de suite. Le fait d'énoncer et de souligner des principes devant guider le développement des technologies est, bien entendu, un travail essentiel. Mais pour les concepteurs, la réflexion ne peut pas s'arrêter là. Ces professionnels doivent aussi articuler les relations entre, d'une part, les principes éthiques proposés et les pratiques organisationnelles ou les choix technologiques, d'autre part[1].

Il y a différentes manières de concevoir la relation entre les valeurs que l'on souhaite mettre de l'avant dans les systèmes d'IA et les mesures concrètes pour y parvenir. Selon la définition de l'approche universaliste, certaines mesures concrètes sont soit nécessaires soit suffisantes dans tous les projets d'IA pour pleinement respecter une valeur précise en jeu. Peu importe les faits de la situation (par exemple : les types d'utilisateurs de la plateforme et l'environnement dans lequel la solution est déployée), certaines mesures concrètes doivent être implémentées pour rendre compte de cette valeur. Les approches contextualistes soutiennent plutôt que différentes mesures concrètes peuvent rendre compte de nos valeurs, et que les actions pertinentes à poser dépendent, dans une large mesure, des faits de la situation.

Pour rendre ces deux thèses plus concrètes, prenons l'exemple de l'avis intitulé *La gestion algorithmique de la main-d'oeuvre : analyse des enjeux éthiques*, publié en 2023 par la CEST. Dans cet avis, la CEST propose notamment d'obliger les employeurs à divulguer le recours à la surveillance électronique au travail, afin de répondre à des

[1] L'absence de lien entre des valeurs et des mesures concrètes peut mener, entre autres, à du fairwashing. Cela s'observe lorsque des organisations affirment qu'elles implémentent des valeurs dans les technologies développées, alors que les mesures concrètes implémentées par l'entreprise n'ont aucun lien avec celles-ci. Voir notamment Aïvodji et al. (2019) sur ce point.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

valeurs éthiques comme la transparence. Deux interprétations peuvent être envisagées pour comprendre la relation entre cette valeur et la mesure recommandée.

Dans une perspective universaliste, la mesure concrète proposée, c'est-à-dire l'obligation de divulguer, pourrait être perçue comme nécessaire (il n'y a pas de transparence sans cette divulgation) ou suffisante (cette divulgation garantit une transparence minimale). Autrement dit, la mesure est une condition (nécessaire ou suffisante) pour atteindre la transparence, indépendamment du contexte. Dans une approche contextualiste, cette recommandation reflète une réponse adaptée aux conditions particulières du Québec en 2023, marquées par des technologies de surveillance électronique et des attentes sociales ou juridiques propres à cette époque. Si le contexte technologique ou sociétal venait à évoluer, par exemple avec l'émergence de nouvelles technologies moins intrusives ou une transformation des attentes éthiques, cette mesure pourrait perdre sa pertinence ou être remplacée par d'autres solutions mieux adaptées. Ces deux interprétations des conclusions de la CEST sont cohérentes avec la démarche de l'organisation, qui vise à articuler des principes éthiques et des recommandations pratiques en fonction des besoins identifiés.

Chercher des mesures concrètes universelles est tentant. Après tout, ces solutions ont le mérite de « durer dans le temps », et de s'appliquer à une foule de situations. Elles permettent d'établir une base commune de mesures nécessaires ou suffisantes pouvant être importées dans toutes les organisations ou tous les systèmes. C'est pourquoi de nombreux programmes de recherche ont pour objectif d'identifier de telles mesures.

Or, est-il réaliste d'espérer trouver des mesures concrètes universelles ? Cette question a récemment fait l'objet de réflexions. Castro et al. (2023), par exemple, proposent une méthodologie de justification d'une mesure concrète, en réponse à un principe éthique dans le domaine de l'IA, soit l'égalité formelle des chances. Pour ce faire, la méthodologie de Castro et al. articule les relations entre quatre éléments, soit :

- Les faits moraux fondamentaux : Ce sont les vérités morales de base qui sous-tendent les jugements éthiques dans une théorie donnée. Par exemple, dans le domaine de l'IA, un fait moral fondamental pourrait être le respect de la vie privée des utilisateurs. Ce principe éthique guide la conception et le déploiement des algorithmes.
- Les faits empiriques pertinents : Ce sont les faits qui ont une influence sur l'évaluation morale dans une situation donnée. En IA, cela pourrait inclure la reconnaissance des biais dans les ensembles de données. Ces biais peuvent entraîner des discriminations involontaires dans les décisions prises par les systèmes d'IA, affectant ainsi leur évaluation éthique.
- Les principes intermédiaires : Ce sont des règles ou des lignes directrices qui permettent d'appliquer des faits moraux fondamentaux à des cas pratiques. Par exemple, un principe intermédiaire en IA pourrait être l'équité et la non-discrimination dans les décisions automatisées.
- Les mesures concrètes : Ce sont des actions ou des pratiques précises recommandées par une théorie éthique dans une situation réelle. Dans le contexte de l'apprentissage automatique, ces mesures pourraient être la présentation d'un formulaire de consentement d'utilisation des données lors du téléchargement d'une application qui utilise l'IA.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

La figure 1 décrit les relations unissant ces quatre notions dans la méthodologie de Castro et al.

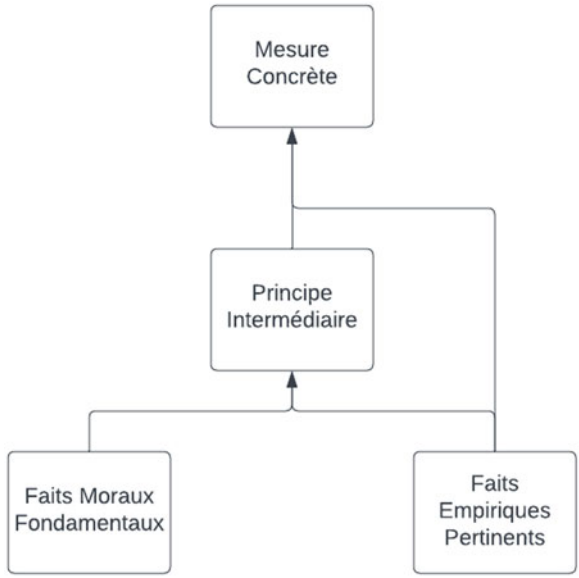

Figure 1. Interactions entre les concepts normatifs introduits par Castro et al. (2023 ; nous traduisons).

Appliquée au principe intermédiaire de l'équité, cette méthodologie amène les auteurs à conclure que les mesures concrètes proposées par les organisations pour atteindre l'égalité des chances dans des algorithmes ne sont ni absolument incontournables ni absolument erronées. Les auteurs concentrent leur attention sur trois mesures concrètes courantes pour améliorer l'équité dans les algorithmes, soit : (i) le traitement anonyme des dossiers des personnes, (ii) l'égalisation des chances entre tous les groupes socioéconomiques et (iii) l'analyse contrefactuelle des informations personnelles. Toutes ces mesures ont été proposées par des institutions privées ou publiques pour respecter le principe d'égalité des chances par des algorithmes. Or, selon les auteurs, aucune de ces mesures n'est absolument nécessaire ou suffisante pour respecter l'équité. L'adéquation des mesures d'égalité des chances dépend du contexte (c'est-à-dire des faits empiriques pertinents) propre à chaque situation. En d'autres termes, lorsqu'appliquée à la notion d'équité, la méthodologie proposée par Castro et al. tend à soutenir une conception contextualiste des mesures concrètes pertinentes[2].

2.2. La méthodologie de Castro et al. appliquée à la XIA

La méthodologie de Castro et al. peut être appliquée aux relations entre principes intermédiaires et mesures concrètes en XIA. En suivant cette même méthodologie,

[2] Les auteurs parlent de « pluralisme » des mesures concrètes.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

nous proposons de poser la transparence comme un fait moral fondamental. La transparence, dans ce contexte, renvoie à la nécessité d'être ouvert et honnête en ce qui concerne les processus, les décisions et les actions. Elle peut être qualifiée de fondamentale, car elle touche à des valeurs morales essentielles comme la confiance, la justice et l'intégrité, des valeurs que Castro et al. identifient comme des points d'ancrage éthiques. Cette approche nous amène à définir la XIA comme principe intermédiaire, c'est-à-dire comme articulation pratique de ce fait moral fondamental dans le domaine technologique. La figure 2 présente notre interprétation de la méthodologie de Castro et al. appliquée à la XIA. Il restera ensuite à savoir si certaines mesures concrètes découlent *universellement* de ces faits moraux et principes, ou si elles doivent être ajustées en fonction des contextes donnés.

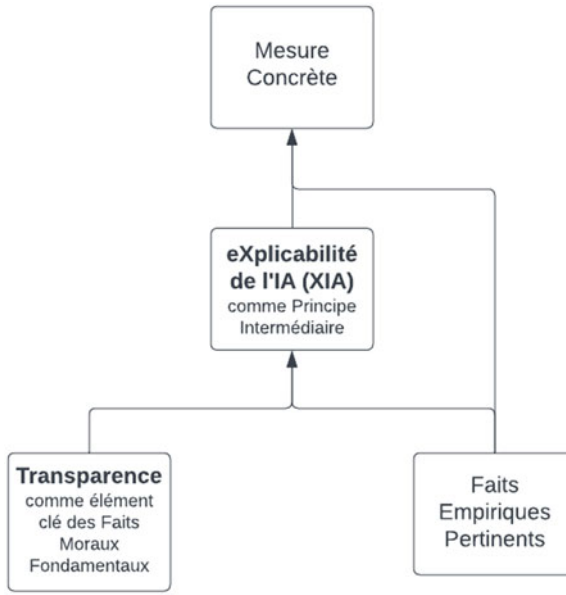

Figure 2. Interprétation de l'application des concepts normatifs introduits par Castro et al. à la XIA.

3. Évaluation normative des mesures concrètes pour le principe intermédiaire de la XIA

Dans cette section, nous souhaitons évaluer le statut (universel ou contextuel) de mesures concrètes de la XIA. Ainsi que le recommandent Castro et al. (2023), cette évaluation sera réalisée en fonction de la conformité de ces mesures avec les faits empiriques pertinents, dans un contexte donné.

Nous mettons en évidence la complexité et la diversité des approches existantes, et démontrons qu'une analyse basée sur les principes intermédiaires et les faits empiriques pertinents aide à choisir la méthode la plus appropriée pour une situation donnée. Cette évaluation pose les deux questions suivantes pour chaque paire de principes intermédiaires et de mesures concrètes :

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

1) Cette mesure concrète est-elle *suffisante* pour satisfaire au principe intermédiaire de la XIA en prenant en considération les faits empiriques pertinents étudiés ?
2) Cette mesure concrète est-elle *nécessaire* pour satisfaire au principe intermédiaire de la XIA en prenant en considération les faits empiriques pertinents étudiés ?

Pour répondre à ces questions, nous évaluons trois mesures concrètes, soit : (i) la publication de notes de transparence IA, (ii) la présentation de l'impact des caractéristiques sur la prédiction à l'utilisateur, et (iii) l'affichage d'un score de similitude entre des données d'entrée. Toutes ces mesures ont été proposées par de grandes institutions ou entreprises pour respecter le principe de la XIA.

3.1. La publication de notes de transparence IA

Nous commençons en examinant une mesure proposée par Microsoft : la note de transparence IA. Chez Microsoft, cette note fait partie d'une initiative plus large qui vise à rendre l'IA compréhensible. Elle fournit des informations textuelles détaillées sur les services IA offerts : le fonctionnement des différents modèles d'IA au sein des services Azure de Microsoft, les choix effectués par les concepteurs de systèmes qui influencent la performance et le comportement des systèmes, ainsi que la façon dont l'ensemble des services IA de Microsoft vont s'intégrer dans le système client, qui inclut la technologie, les personnes et l'environnement (Liao et al., 2023). Pour Azure AI Language, la note de transparence décrit les utilisations possibles des différentes fonctionnalités, telles que la reconnaissance d'entités nommées, la détection de langues, l'analyse de sentiments, et bien d'autres (Aahill et al., 2023). Chaque fonctionnalité est accompagnée de cas d'usage, de considérations sur les limites potentielles et de conseils pour améliorer la performance des systèmes. Cela inclut également des informations sur la gestion de la confidentialité des données et la personnalisation de la classification des textes.

Si l'on prenait le point de vue des autorités de régulation, il semblerait logique de s'assurer que les entreprises se conforment bien à la publication d'une note de transparence pour augmenter la compréhension d'une IA. Sur la base de cette observation, la présentation d'une note de transparence paraît nécessaire pour répondre aux attentes de la XIA.

En revanche, notons que Microsoft n'affiche pas une note de transparence pour tous les systèmes d'IA mis à la disposition de ses clients. Par exemple, l'entreprise présente une fonctionnalité de vérification des pourriels qui fait appel à l'IA pour analyser le contenu de chaque message de marketing par courrier électronique (Ferguson et al., 2024). L'IA génère une probabilité que le courrier soit un pourriel. Il n'y a pas de note de transparence pour cette fonctionnalité. Les prédictions émises par l'IA ont un faible impact sur le quotidien des utilisateurs. Les utilisateurs comprennent généralement que le système vise à améliorer l'efficacité, sans pour autant affecter significativement leur travail. Aussi, dans ce scénario, l'application de l'IA est relativement simple et bien comprise. Dans ce cas, la note de transparence est

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

considérée comme non nécessaire. Cela indique que la note de transparence est nécessaire pour *certains* produits, mais pas pour *tous* les produits.

Lorsqu'on se penche sur la question de la suffisance, il est assez facile de voir que l'utilisation unique de la note de transparence pour tous les scénarios est loin d'être suffisante. Ce n'est pas parce que l'on connaît les principes généraux d'une IA que l'on comprend comment ceux-ci se manifestent dans un cas particulier, tel que le refus d'un prêt hypothécaire à un client.

3.2. La présentation de l'impact des caractéristiques sur la prédiction à l'utilisateur

Examinons maintenant une autre mesure proposée en lien avec la XIA, soit la présentation de l'impact des caractéristiques d'entrée de l'algorithme sur la prédiction. L'application mobile Intact Assurance permet aux utilisateurs d'obtenir des rabais sur leur assurance automobile basés sur une analyse intelligente de leurs habitudes de conduite (Intact Assurance, s. d.). Cette application leur permet aussi d'observer ces habitudes et l'impact qu'elles ont sur le rabais qui leur est offert, un exemple concret de la présentation de l'impact des caractéristiques d'entrée de l'algorithme sur la prédiction.

Tout comme la mesure précédente, s'il nous était demandé, en tant qu'autorité de réglementation, de veiller à ce que les entreprises respectent effectivement l'idéal de la XIA, il semblerait raisonnable de vérifier que les systèmes fournissent des explications sur l'impact des caractéristiques d'entrée d'une IA. Pour des produits comme une police d'assurance, où les décisions de l'IA peuvent avoir un impact monétaire direct sur les utilisateurs, la compréhension de ces impacts devient cruciale pour obtenir la meilleure prime possible. Cette description de l'importance des caractéristiques serait donc jugée nécessaire pour répondre aux exigences de la XIA, ne serait-ce que pour certains produits comme une police d'assurance.

Or, cette mesure concrète est-elle nécessaire ou suffisante pour tous les produits ? Certains faits de la situation pourraient affecter notre évaluation de la satisfaction de cette mesure. Pour illustrer nos propos, prenons l'exemple du diagnostic du cancer par imagerie médicale, un système que l'on dit couramment de haute incidence.

Dans ce contexte, la nécessité de cette mesure dépend de l'objectif et de l'utilisateur visé. Pour un expert en IA ou un professionnel de la santé, la présentation de l'impact de chaque pixel de l'image pourrait être jugée nécessaire pour vérifier le fonctionnement correct du système et valider son raisonnement. En revanche, pour un patient, ces informations détaillées sont non nécessaires, car elles ne comblent pas directement son besoin : comprendre pourquoi le diagnostic a été posé et ce que ça implique pour son traitement ou sa santé. La complexité des données, comme l'analyse de milliers de pixels, rend l'information inutilisable pour un non-expert. Le patient aurait plutôt besoin d'explications simplifiées, comme une description textuelle indiquant les principales caractéristiques qui ont influencé la décision (par exemple, la taille de la tumeur ou son emplacement), ou une visualisation intuitive mettant en évidence les zones concernées sur l'image. Ainsi, la présentation de l'impact des caractéristiques d'entrée (les pixels) sur la prédiction n'est pas une mesure nécessaire pour satisfaire les attentes de la XIA dans tous les cas d'utilisation.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

Lorsque l'on se penche sur la question de la suffisance, la réponse peut être nuancée. Pour un médecin, bien que cette mesure puisse fournir des informations utiles pour analyser les décisions de l'IA, elle n'est pas suffisante à elle seule. Les médecins ont souvent besoin d'une synthèse des facteurs influents qui relie directement les données techniques aux implications cliniques, tels que la taille de la tumeur ou ses caractéristiques biologiques. Ainsi, si cette mesure peut suffire dans un cadre purement technique (par exemple, pour un expert en IA), elle doit souvent être accompagnée d'une interprétation clinique pour être pleinement exploitable dans un contexte médical.

3.3. L'affichage d'un score de similitude entre des données d'entrée

Finalement, examinons une dernière mesure concrète : l'affichage d'un score de similitude entre des données d'entrée. Pour comprendre de quoi il s'agit, pensons, par exemple, à un système de recommandation de films, tel que celui employé par Netflix, qui présente un score (pourcentage) de similitude entre deux films (Netflix, s. d.).

Dans ce cas, l'utilisateur interagit avec le système en choisissant des films, et le système utilise ces informations pour recommander d'autres titres similaires. Les utilisateurs bénéficient de ces recommandations personnalisées sans avoir besoin de comprendre les détails techniques sur la façon dont les recommandations sont générées. Ils sont généralement plus intéressés par la pertinence des recommandations que par les mécanismes sous-jacents du système de recommandation. Dans ce cas, nous pourrions dire que, dans le contexte précis du choix d'un film le samedi soir, l'affichage d'un score de similitude est une mesure suffisante en ce sens qu'elle excède ce qui est requis pour permettre aux abonnés de comprendre le fonctionnement de la plateforme. Or, cette mesure n'est pas nécessaire. D'autres moyens que le score de similitude pourraient favoriser cette compréhension. Par exemple, la plateforme pourrait expliquer ses recommandations au moyen des préférences passées des utilisateurs : « Nous vous recommandons le film X car vous avez aimé le film Y. » Une telle justification permet aux utilisateurs de comprendre les recommandations, sans qu'aucun score soit mobilisé.

Examinons l'application de cette mesure concrète à d'autres contextes. Imaginons, par exemple, qu'un médecin fasse un diagnostic en présentant seulement le score de similitude entre deux patients généré par une IA. En d'autres termes, le médecin se contente de dire que le patient X a sans doute la maladie M, puisqu'il a des caractéristiques semblables à un autre patient Y aux prises avec la maladie M. Cette explication hautement limitée ne respecte pas le critère de la XIA dans le secteur médical. Dans ce cas, la mesure est nécessaire en ceci qu'elle fournit une base pour établir un diagnostic et soutient la démarche explicative du médecin. Cependant, cette mesure est insuffisante pour répondre pleinement aux attentes de la XIA, car elle ne permet pas de comprendre les raisons sous-jacentes de cette similitude. Elle ne répond pas aux questions essentielles, telles que : quelles caractéristiques précises du patient X ont contribué à cette prédiction ? Pourquoi ces caractéristiques sont-elles particulièrement importantes ?

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

3.4. Conclusion partielle

Faisons le point. Nous avons étudié trois mesures concrètes qui ont été proposées en lien avec la XIA, soit : (i) la publication de notes de transparence IA, (ii) la présentation de l'impact des caractéristiques sur la prédiction à l'utilisateur, et (iii) l'affichage d'un score de similitude entre des données d'entrée. Et nous avons conclu que celles-ci n'étaient pas absolument nécessaires ni suffisantes pour satisfaire au principe intermédiaire de la XIA dans tous les contextes.

Ces résultats mettent en évidence le caractère non universel des mesures étudiées. Leur pertinence dépend des spécificités contextuelles, telles que le type d'utilisateur, le domaine d'application ou les objectifs visés par la technologie proposée. En conséquence, les mesures concrètes possèdent un statut contextuel qui appelle une adaptation au cas par cas pour répondre aux exigences de la XIA.

4. Proposition d'une matrice normative pour l'évaluation des mesures en XIA

En suivant l'analyse réalisée à la section 3, l'un des principaux défis dans l'évaluation normative de la pertinence d'une mesure concrète pour la XIA réside dans l'identification correcte des faits empiriques pertinents. Cette analyse justifie les recherches en XIA se spécialisant dans un contexte précis.

Mais ce constat soulève un problème. Conformément à une approche contextualiste, devons-nous simplement nous dire que « tout dépend de la situation donnée » ? Pour des ingénieurs en IA qui souhaitent satisfaire à des principes éthiques intermédiaires dans leurs travaux, cette conclusion n'est d'aucun secours. Elle compromet aussi les collaborations entre des entreprises de différents secteurs désirant partager leurs pratiques. Une entreprise a beau avoir développé des pratiques satisfaisantes pour atteindre l'explicabilité dans son domaine, cela ne signifie nullement que d'autres entreprises devraient (ou même pourraient) adopter les mêmes pratiques. Ainsi, le défi consiste à rendre compte du caractère contextuel des mesures concrètes pertinentes, tout en offrant de l'accompagnement aux ingénieurs souhaitant intégrer l'éthique à leur pratique professionnelle.

Pour surmonter les difficultés que les ingénieurs rencontrent dans l'évaluation éthique de la pertinence des mesures concrètes, il est essentiel de les orienter vers une méthode permettant de préciser, relativement aux faits de la situation, lesquelles sont pertinentes. Une telle approche favorise la compréhension des enjeux et encourage une prise de décision éclairée et adaptée au contexte précis de chaque projet d'IA.

Pour progresser dans la compréhension et l'application de mesures concrètes qui soutiennent le principe intermédiaire de la XIA, nous proposons la matrice d'évaluation normative des mesures concrètes de la XIA. Cette matrice est qualifiée d'expérimentale puisqu'elle a été validée à partir de cas d'utilisation fictifs. Cette approche consiste en la création d'une matrice, qui prend en considération deux faits empiriques pertinents et leurs variations possibles. Chaque axe de la matrice représente un fait empirique différent, tandis que les cellules formées à l'intersection de ces axes reflètent les combinaisons possibles de variations de ces faits. En fonction de la position exacte d'un système dans cette matrice, une entreprise peut déterminer de manière proactive les mesures concrètes les mieux adaptées à son contexte. Le tableau 1 présente le gabarit d'une telle matrice.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

Tableau 1 : Gabarit de la matrice d'évaluation normative des mesures concrètes de la XIA

		Fait empirique pertinent 2		
		Variation 1	Variation 2	Variation 3
Fait empirique pertinent 1	Variation 1	Mesure concrète 1.1	Mesure concrète 1.2	Mesure concrète 1.3
	Variation 2	Mesure concrète 2.1	Mesure concrète 2.2	Mesure concrète 2.3
	Variation 3	Mesure concrète 3.1	Mesure concrète 3.2	Mesure concrète 3.3

En intégrant explicitement les faits empiriques pertinents dans la matrice, cette approche offre un outil flexible qui s'adapte aux spécificités de chaque contexte d'application. Par exemple, dans un système d'IA déployé dans le secteur médical, les solutions pourront être personnalisées en fonction de facteurs tels que la complexité des données cliniques, les besoins précis des utilisateurs (médecins, patients), ou encore l'emplacement physique de déploiement du système (urgence, centre de radiologie).

De plus, ce cadre offre un moyen structuré d'explorer les interactions entre les faits empiriques pertinents. En identifiant les combinaisons de ces facteurs dans la matrice, il devient possible d'évaluer comment les variations d'un facteur modifient l'efficacité ou la pertinence d'une mesure concrète. Cela encourage une réflexion normative proactive, car les utilisateurs de la matrice peuvent anticiper des tensions ou des compromis éthiques liés à ces interactions, et ainsi adapter leurs choix en conséquence.

Ce cadre permet non seulement une personnalisation des solutions de XIA, mais encourage également une réflexion normative sur la manière dont divers facteurs empiriques interagissent et influencent ces solutions.

5. Étude de cas dans le secteur aéronautique

Pour illustrer cette approche, considérons le cas d'un constructeur d'avions. Un avion intègre de multiples systèmes aux fonctions variées, lesquels interagissent de différentes manières avec plusieurs utilisateurs. Afin d'élaborer une matrice de mesures concrètes pour la XIA, le constructeur choisit de se concentrer sur deux faits empiriques pertinents. Le premier est l'expertise de l'utilisateur du système dans le domaine aéronautique. L'utilisateur est défini en fonction du système analysé (par exemple : mécanicien, passager ou pilote) et un niveau de connaissance en IA lui est attribué avec trois variations possibles, soit basique, moyen ou élevé. Le deuxième fait empirique pertinent de cette matrice est le niveau de risque du système pour les passagers d'un avion, qui présente aussi trois variations possibles, soit minimal, limité ou élevé. Ainsi, à la suite d'une évaluation normative des faits empiriques pertinents et des mesures concrètes réalistes et accessibles, le constructeur serait amené à créer une matrice semblable à celle présentée au tableau 2. Pour ce faire, le constructeur aura défini, pour chaque combinaison de variations possible, une ou plusieurs mesures concrètes. Pour valider l'efficacité de cette matrice, évaluons si les mesures concrètes qui y sont présentées sont nécessaires et suffisantes pour deux systèmes d'IA que l'on pourrait retrouver à bord d'un avion. Notons que la matrice suggérée au tableau 2 ne sert pas de référence pour les constructeurs d'avions, mais plutôt d'exemple pour guider le format et l'applicabilité d'un tel outil.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

Tableau 2 : Matrice de mesures concrètes de la XIA en fonction de faits empiriques pertinents

		Expertise dans le domaine aéronautique de l'utilisateur du système		
		Basse	Moyenne	Élevée
Niveau de risque pour les passagers	Risque minimal	Démontrer le fonctionnement correct du système au moyen de métriques de performance universellement connues.	Démontrer le fonctionnement correct du système au moyen de métriques de performance connues du secteur d'activité global de l'aéronautique.	Démontrer le fonctionnement correct du système au moyen de métriques de performance basées sur des informations techniques approfondies sur le secteur d'activité donné.
	Risque limité	Présenter les grandes catégories de données d'entrée du système en fournissant un lexique qui explique l'utilité de chacune de ses catégories.	Présenter partiellement les données d'entrée du système — celles qui sont généralement connues dans le domaine de l'aéronautique.	Présenter les données d'entrée qui ont mené aux prédictions, incluant des informations techniques approfondies sur le secteur d'activité donné.
	Risque élevé	Fournir des scénarios « et si » de complexité moindre, expliquant ce qui pourrait se passer si différentes actions sont entreprises. Ajouter une description textuelle et visuelle.	Fournir des scénarios « et si » en temps réel des cas les plus fréquents, expliquant ce qui pourrait se passer si différentes actions sont entreprises.	Offrir à l'utilisateur l'option d'élaborer des scénarios « et si », expliquant ce qui pourrait se passer si différentes actions sont entreprises.

Le premier cas de figure que nous souhaitons explorer est le développement d'un système de gestion optimisée de la consommation de carburant. Ce système utilise des algorithmes d'apprentissage automatique pour optimiser la consommation de carburant de l'avion en calculant les trajectoires de vol les plus efficaces en temps réel, prenant en compte les conditions météorologiques actuelles, le poids de l'avion, et d'autres variables. Cela peut réduire les coûts d'exploitation pour les compagnies aériennes et diminuer l'empreinte carbone des vols. Ce système n'affecte pas directement la sécurité des passagers, mais améliore l'efficacité et la durabilité des opérations aériennes. Les passagers bénéficient indirectement de ces améliorations par des coûts potentiellement réduits et une conscience accrue de l'empreinte environnementale de leur voyage. Les ingénieurs chargés du développement de ce système pourraient déterminer que celui-ci présente un niveau de risque minimal pour les passagers et que l'expertise de ces derniers en matière aéronautique est limitée. En suivant la matrice des mesures concrètes de la XIA en fonction des faits empiriques pertinents (tableau 2), qui aurait préalablement été construite par la compagnie d'aviation, une mesure concrète de la XIA consisterait à démontrer le fonctionnement correct du système au moyen de métriques de performance universellement connues. Cela pourrait se traduire, par exemple, par l'affichage en temps réel, sur l'écran des passagers, des économies de carburant réalisées grâce au système durant le vol.

Bien que l'expertise en aéronautique des passagers soit limitée, fournir des informations compréhensibles et directement liées à l'efficacité du système favorise

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

l'appréciation de la technologie et de ses avantages environnementaux. Cela peut également contribuer à une prise de conscience plus large des efforts de durabilité dans le secteur aérien. En ce sens, la mesure est nécessaire pour renforcer la confiance et la compréhension publiques quant à l'utilisation des technologies d'IA dans l'amélioration des opérations aériennes. De plus, l'intérêt principal des passagers envers ce système concerne l'efficacité et les implications environnementales des vols qu'ils font. Ainsi, la présentation des économies de carburant réalisées grâce au système peut être considérée comme une mesure suffisante de la XIA. Cette information répond directement à la préoccupation des passagers concernant la durabilité et l'impact écologique de leurs voyages, en fournissant une mesure tangible des efforts de la compagnie aérienne pour minimiser son empreinte carbone.

Un deuxième cas de figure illustrant notre approche serait le développement d'un système de maintenance prédictive pour les avions. Ce système d'IA analyse les données issues des capteurs installés sur diverses composantes de l'avion, telles que les moteurs, les systèmes hydrauliques et les équipements électriques, pour prédire les pannes potentielles avant qu'elles ne se produisent. En s'appuyant sur l'apprentissage automatique et l'analyse prédictive, le système identifie les signes avant-coureurs de défaillance, permettant aux équipes d'entretien d'intervenir de manière proactive, plutôt que réactive. Cela aide à éviter les retards et les annulations de vols dus à des problèmes techniques inattendus, améliorant ainsi l'expérience globale des passagers et la fiabilité des opérations aériennes. Bien que ce système ne soit pas directement lié à la sécurité des vols en temps réel, il joue un rôle crucial dans la prévention des incidents et assure le bon fonctionnement de l'avion. La maintenance prédictive contribue à une meilleure gestion des ressources et à une réduction des coûts pour les compagnies aériennes, tout en augmentant la satisfaction des passagers grâce à une diminution des perturbations de voyage. Les ingénieurs chargés du développement de ce système pourraient déterminer que celui-ci présente un niveau de risque limité pour les passagers et que l'expertise des utilisateurs du système (techniciens et ingénieurs mécaniques) en aéronautique est élevée. En se référant à la matrice définie au tableau 2, une mesure concrète de la XIA pourrait être de présenter les données d'entrée qui ont mené à ces prédictions, incluant des informations techniques approfondies, adaptées au niveau d'expertise élevé des utilisateurs. Ceci pourrait être réalisé au moyen de rapports fournis aux utilisateurs du système.

La présentation des données d'entrée qui ont conduit aux prédictions du système, incluant des informations techniques approfondies, est indéniablement nécessaire pour les mécaniciens. Cette nécessité découle du rôle clé de ces derniers dans la prévention des défaillances et dans le maintien de la sécurité et de la fiabilité des appareils. L'accès à des informations détaillées leur permet de comprendre les fondements des alertes de la maintenance prédictive, facilitant ainsi l'identification rapide et précise des problèmes potentiels et la planification des interventions. Pour les mécaniciens, la présentation des données d'entrée ayant mené aux prédictions du système est potentiellement suffisante *dans leur contexte*. Ceux-ci sont des experts dans le domaine aéronautique et sont familiarisés avec les nuances techniques complexes des systèmes aériens. Donc, l'évaluation de la suffisance prend en compte leur haut niveau de compétence et leur capacité à interpréter des données complexes pour diagnostiquer et prévenir efficacement les problèmes techniques.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

Par le biais de ces exemples, nous pouvons valider la pertinence de la matrice des mesures concrètes de la XIA en fonction des faits empiriques pertinents. Elle permet aux ingénieurs d'avoir des lignes directrices établies et approuvées par l'entreprise, élaborées en tenant compte des faits empiriques jugés pertinents pour le secteur d'activité concerné. Elle permet aussi aux ingénieurs de différentes organisations d'évaluer la pertinence d'adopter des mesures communes et de partager leur expertise. Il leur suffit de déterminer si les faits pertinents de leurs organisations respectives sont suffisamment similaires.

Il convient d'établir que plusieurs améliorations doivent être envisagées avant d'industrialiser cette approche. Par exemple, plusieurs autres faits pertinents pourraient être pris en considération, tels que l'expertise en IA des utilisateurs ou encore le temps de compréhension accessible à l'utilisateur du système. Dans le cas d'un système d'assistance de pilotage pour des situations d'urgence, le pilote n'aurait que quelques secondes pour comprendre les explications qui lui sont fournies, alors que dans le cas du système de maintenance prédictive pour les avions, le mécanicien dispose de plus de temps. Des matrices supplémentaires ou à dimensions plus grandes devraient être explorées. En outre, la réussite de l'approche dépend fortement de la volonté et de la capacité des organisations d'investir dans des évaluations normatives et d'adapter leurs pratiques. Dans certains cas, la pression pour une mise sur le marché rapide pourrait les pousser à négliger ces considérations éthiques au profit de l'efficacité opérationnelle.

Conclusion

Dans cet article, nous avons exploré les dimensions normatives de la XIA, avec un accent particulier sur les défis que les ingénieurs IA rencontrent dans l'intégration de mesures concrètes. Pour ce faire, notre étude a abordé les deux questions suivantes :

1) Les mesures de la XIA peuvent-elles être standardisées universellement, ou doivent-elles être adaptées spécifiquement à chaque contexte d'application ?
2) Quelles méthodes et quels outils peuvent aider les ingénieurs à intégrer efficacement des considérations éthiques de la XIA dans le développement de l'IA ?

Notre recherche a d'abord mis en évidence que le choix d'une mesure de la XIA plutôt qu'une autre doit être adapté à chaque contexte pour être véritablement efficace. Cette conclusion est soutenue par l'analyse de différentes applications de la XIA dans des contextes industriels variés, où les exigences et les impacts de l'IA diffèrent grandement. Bien que des standards universels puissent fournir un cadre général, une personnalisation de ces mesures est nécessaire pour répondre précisément aux besoins éthiques, techniques et opérationnels de chaque projet. En soi, documenter la pertinence de certaines pratiques dans un contexte donné représente une avancée dans la discussion collective sur la XIA.

Pour aider les ingénieurs à intégrer efficacement l'explicabilité dans le développement de l'IA, notre recherche a proposé et détaillé la mise en oeuvre d'une matrice d'évaluation normative des mesures concrètes de la XIA. Cette matrice, qui devrait être produite en amont du développement des systèmes d'IA, guide les ingénieurs à travers un processus structuré pour définir les mesures concrètes en

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

réponse aux principes éthiques dans des situations variées. En fournissant un cadre méthodologique, la matrice permet d'augmenter la conformité aux normes éthiques et d'adapter les mesures concrètes de la XIA au contexte spécifique de chaque application. Cette approche pose les bases de l'établissement d'une méthode d'évaluation éthique des mesures de la XIA pour les ingénieurs.

Précisons finalement que la matrice d'évaluation normative pourrait être utilisée à d'autres fins. Elle pourrait notamment être employée par des institutions publiques pour clarifier la portée de leurs recommandations. Comme nous l'avons mentionné dans la section 2, des institutions publiques comme la CEST proposent aux acteurs gouvernementaux des mesures concrètes visant à respecter différentes valeurs, allant de la transparence à la sécurité, en passant par le bien-être. Or, les avis de la CEST ne spécifient pas si les mesures recommandées sont relatives au contexte étudié, ou si elles sont universelles (les deux interprétations sont cohérentes avec la démarche de la CEST). À supposer que certaines des mesures proposées par la CEST soient relatives au contexte, le recours aux matrices d'évaluation permettrait à cette institution de mieux souligner l'adéquation entre (i) ses recommandations et (ii) les faits empiriques pertinents qu'elle prend en compte. D'une certaine façon, on pourrait dire que la matrice d'évaluation normative est un outil pour transmettre plus d'information aux décideurs quant aux éléments contextuels justifiant des mesures concrètes.

Remerciements. Nous remercions le Laboratoire d'ingénierie cognitive et sémantique (LiNCS) de l'École de technologie supérieure (ÉTS) pour son soutien, le Groupe de recherche interuniversitaire sur la normativité (GRIN) et le Centre de recherche informatique de Montréal (CRIM). Nous remercions également les évaluateurs anonymes pour leurs commentaires constructifs.

Conflits d'intérêts. Les auteurs déclarent qu'ils n'ont pas de conflit d'intérêt pour cette recherche.

Références bibliographiques

Aahill, Nitinme, Urban, E. et P. Faley (2023). Transparency Note for Azure AI Language. *Microsoft Legal Resources*. https://learn.microsoft.com/en-us/legal/cognitive-services/language-service/transparency-note/

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S. et A. Tapp (2019). Fairwashing: The Risk of Rationalization. Dans K. Chaudhuri et R. Salakhutdinov (dir.), *Proceedings of the 36th International Conference on Machine Learning* (p. 161-170). PMLR.

Anders, C. J., Neumann, D., Samek, W., Müller, K. R. et S. Lapuschkin (2021). Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy. *arXiv*.

Belaid, M. K., Hüllermeier, E., Rabus, M. et R. Krestel (2022). Do We Need Another Explainable AI Method? Toward Unifying Post-hoc XAI Evaluation Methods into an Interactive and Multi-dimensional Benchmark. *arXiv*.

Benchekroun, O., Rahimi, A., Zhang, Q. et T. Kodliuk (2020). The Need for Standardized Explainability. *arXiv*.

Castro, C., O'Brien, D. et B. Schwan (2023). Egalitarian Machine Learning. *Res Publica*, *29*(2), 237–264. https://doi.org/10.1007/s11158-022-09561-4

Commission de l'éthique en science et en technologie (2021). *Les effets de l'intelligence artificielle sur le monde du travail et la justice sociale*. Gouvernement du Québec.

Commission de l'éthique en science et en technologie (2023a). *La gestion algorithmique de la main-d'oeuvre : analyse des enjeux éthiques*. Gouvernement du Québec.

Commission de l'éthique en science et en technologie (2023b). *La transformation numérique du réseau de la santé et des services sociaux en vue d'intégrer l'intelligence artificielle : un regard éthique*. Gouvernement du Québec.

https://doi.org/10.1017/S0012217325000095 Published online by Cambridge University Press

Commission de l'éthique en science et en technologie (2023c). *Mériter et renforcer la confiance des citoyens dans la gestion et la valorisation des données de santé : pour une gouvernance transparente et responsable, soucieuse de la dignité des personnes et de l'intérêt public*. Gouvernement du Québec.

Commission de l'éthique en science et en technologie (2024). *Intelligence artificielle générative en enseignement supérieur : enjeux pédagogiques et éthiques*. Gouvernement du Québec.

Commission européenne (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri = celex%3A52021PC0206

Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies (2019). *Lignes directrices en matière d'éthique pour une IA digne de confiance*. Office des publications de l'Union européenne. https://data.europa.eu/doi/10.2759/74304

Felzmann, H., Fosch-Villaronga, E., Lutz, C. et A. Tamò-Larrieux (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Ferguson, A., Munjal, A et M. Hartmann (2024). Use AI to Check Your Message-Content Spam Score. *Microsoft Customer Insights*. https://learn.microsoft.com/en-us/dynamics365/customer-insights/journeys/spam-checker

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme*, *19*(4), 534–560. https://doi.org/10.1017/epi.2022.39

Garrett, N., Beard, N. et C. Fiesler (2020). More Than "If Time Allows": The Role of Ethics in AI Education. Dans A. Markham, J. Powles, T. Walsh et A. L. Washington (dir.), *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (p. 272-278). Association for Computing Machinery.

Intact Assurance (s. d.). Termes et conditions de l'application Mon Intact. https://www.intact.ca/fr/termes-application-intact#maconduite-v4

Le, P. Q., Nauta, M., Nguyen, V. B., Pathak, S., Schlötterer, J. et C. Seifert (2023). Benchmarking EXplainable AI: A Survey on Available Toolkits and Open Challenges. *International Joint Conferences on Artificial Intelligence*.

Liao, Q. V., Subramonyam, H., Wang, J. et J. Wortman Vaughan (2023). Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-powered User Experience. Dans *Proceedings of the 2023 CHI conference on human factors in computing systems* (p. 1-21).

Liao, Q. V. et K. R. Varshney (2021). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv*.

Memarian, B. et T. Doleck (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and Higher Education: A Systematic Review. *Computers and Education: Artificial Intelligence*, *5*, 100–152. https://doi.org/10.1016/j.caeai.2023.100152

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M. et C. Seifert (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42. https://doi.org/10.1145/3583558

Netflix (s. d.). How to Rate TV Shows and Movies? *Netflix Help Center*. https://help.netflix.com/en/node/9898

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Nyrup, R. et D. Robinson (2022). Explanatory Pragmatism: A Context-Sensitive Framework for Explainable Medical AI. *Ethics and Information Technology*, *24*(1). https://doi.org/10.1007/s10676-022-09632-3

Parthasarathy, V., Urban, E. et P. Faley (2023). Transparency Note for Image Analysis. *Microsoft Legal Resources*. https://learn.microsoft.com/en-us/legal/cognitive-services/computer-vision/imageanalysis-transparency-note

Reddy, G. P. et Y. V. P. Kumar (2023). Explainable AI (XAI): Explained. Dans *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)* (p. 1-6). IEEE.

Rubel, A., Castro, C. et A. Pham (2021). *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press.

van Wynsberghe, A. (2016). *Healthcare Robots: Ethics, Design and Implementation*. Routledge.

van Berkel, N., Tag, B., Goncalves, J. et S. Hosio (2022). Human-Centred Artificial Intelligence: A Contextual Morality Perspective. *Behaviour & Information Technology*, *41*(3), 502–518. https://doi.org/10.1080/0144929x.2020.1818828

Weber, L., Lapuschkin, S., Binder, A. et W. Samek (2023). Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement. *Information Fusion*, *92*, 154–176. https://doi.org/10.1016/j.inffus.2022.11.013

Zhou, J., Gandomi, A. H., Chen, F. et A. Holzinger (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593

Comment citer cet article: Raymond, C., Daoust, M-K., & Ratté, S. (2025). Le développement d'une IA explicable : entre principes éthiques généraux et mesures concrètes. *Dialogue* 64(1), 81–99. https://doi.org/10.1017/S0012217325000095