L'intelligence artificielle peut-elle être une innovation responsable ?

Louis Vuarin, Pedro Gomes Lopes, David Massé

Dans Innovations 2023/3 N° 72 , pages 103 à 147
Éditions De Boeck Supérieur

ISSN 1267-4982
ISBN 9782807380189
DOI 10.3917/inno.pr2.0153
Date de mise en ligne : 06/10/2023

L'intelligence artificielle peut-elle être une innovation responsable ?

Louis VUARIN
Institut interdisciplinaire de l'innovation (i3) – Sciences Économiques et Sociales (SES) Telecom Paris, CNRS, Institut polytechnique de Paris, Palaiseau, France
louis.vuarin@telecom-paris.fr

Pedro GOMES LOPES
Institut interdisciplinaire de l'innovation (i3) – Centre de recherche en gestion (CRG), École polytechnique, CNRS, Institut polytechnique de Paris, Palaiseau, France
pedro.gomes-lopes@polytechnique.edu

David MASSÉ
Institut interdisciplinaire de l'innovation (i3) – Sciences Économiques et Sociales (SES) Telecom Paris, CNRS, Institut polytechnique de Paris, Palaiseau, France
david.masse@telecom-paris.fr

RÉSUMÉ
L'intelligence artificielle (IA) représente un défi majeur pour l'innovation responsable (IR) en raison de l'ampleur des transformations sociétales et productives qu'elle induit. L'enjeu de l'inscription de l'IA dans l'innovation responsable ne réside pas tant dans l'absence de concepts théoriques pour guider son développement, mais plutôt dans la multiplication de ces concepts (IA explicable, IA transparente, IA durable, IA juste…) et dans leur discutable propension à se traduire en actions concrètes réellement transformatrices des écosystèmes d'innovations. À partir d'une analyse bibliométrique de la littérature, cet article propose une cartographie des principaux concepts contribuant à inscrire l'IA dans une démarche d'innovation responsable. Les résultats mettent en lumière la difficile articulation des concepts entre eux, notamment au regard de la concurrence entre les acteurs et les dispositifs d'action préconisés. Cette étude contribue à la littérature sur les défis de l'adoption de l'IA et de son inscription dans une

 DOI: 10.3917/inno.072.0101

démarche d'innovation responsable, elle discute également des risques et opportunités associés à la multiplication de concepts pour inscrire des technologies émergentes dans une démarche d'innovation responsable.

MOTS CLÉS : Intelligence artificielle, Innovation responsable, Performativité, Technologies émergentes

CODES JEL : O32, O33

ABSTRACT

Linking Artificial Intelligence to Responsible Innovation

Artificial intelligence (AI) represents a major challenge for responsible innovation because of the scale of the societal and productive transformations it brings about. The challenge of including AI in responsible innovation does not lie so much in the absence of theoretical concepts to guide its development, but rather in the high number of these concepts (explainable AI, transparent AI, sustainable AI, fair AI…) and in their questionable propensity to be translated into concrete actions that truly transform innovation ecosystems. Based on a systematic review of the literature, this article analyzes the main concepts contributing to the inclusion of AI in a responsible innovation approach. The results highlight the difficulty of linking the concepts together, particularly given the competition between the actors who may prioritize different concepts. This study contributes to the literature on the challenges of adopting AI and including it in a responsible innovation approach. It also discusses the risks and opportunities associated with the multiplication of concepts for including emerging technologies in a responsible innovation approach.

KEYWORDS: Artificial Intelligence, Responsible Innovation, Performativity, Emerging Technologies

JEL CODES: O32, O33

L'intelligence artificielle (IA) représente un défi pour l'innovation responsable (IR) (Brundage, 2016 ; Stahl, Wright, 2018 ; Buhmann, Fieseler, 2021, 2022 ; Stahl, 2022). L'ampleur des transformations sociétales et productives induites (l'impact sur l'emploi, la protection de la vie privée, les biais et la discrimination, la consommation énergétique…) par cette technologie dans un futur proche (Makridakis, 2017 ; Kaplan, Heinlein, 2019) représente un test pour l'innovation responsable, et sa capacité à influencer le développement des technologies émergentes à fort impact. Or, l'IA présente une particularité majeure : elle se caractérise par un développement polycentrique qui constitue une zone d'ombre

pour l'innovation responsable (Stahl, 2022). Il résulte de ce développement polycentrique un fourmillement d'initiatives et de concepts visant à renforcer la dimension éthique du développement et du déploiement de cette technologie (Mittelstadt *et al.*, 2016 ; Brundage, 2016). Le défi de l'inscription de l'IA ne réside alors pas tant dans l'absence de concepts théoriques pour guider son développement, mais plutôt dans la multiplication de ces concepts, et dans leur discutable propension à se traduire en actions concrètes réellement transformatrices des écosystèmes d'innovations (Brundage, 2016 ; Stahl, 2022).

De nombreux auteurs (Mittelstadt *et al.*, 2016 ; Mittelstadt, 2019 ; Santoni de Sio, Mecacci 2021 ; Merhi, 2022), confortés par des études de terrain (Vakkuri *et al.*, 2022), mettent en effet en garde contre des concepts qui resteraient inopérants en l'absence d'acteurs dédiés à leur mise en œuvre et de dispositifs efficaces associés. Autrement dit, c'est leur caractère *performatif* qui est remis en question – c'est-à-dire, leur potentiel à fédérer des réseaux d'acteurs et à diffuser des dispositifs d'action qui entrent en synergie au point de changer la réalité pratique des organisations et des marchés (Cabantous, Gond, 2011 ; Aggeri, 2017). Dans cette optique, l'objectif de cet article est d'apporter des éléments de réponse aux interrogations suivantes : Quels sont les principaux concepts qui contribuent à inscrire l'IA dans une démarche d'innovation responsable ? Comment caractériser la performativité de ces concepts ?

L'article présente tout d'abord une cartographie des concepts clés qui ont été identifiés à partir d'une analyse bibliométrique de la littérature portant sur la relation entre l'IA et l'innovation responsable. Cette analyse se base sur un corpus de 207 articles, et a utilisé une méthode de *clusterisation* pour identifier les co-occurrences thématiques. Pour chaque concept identifié, l'article présente son fondement théorique, les dispositifs qui lui sont associés, ainsi que les réseaux d'acteurs qui le promeuvent. L'article évalue ensuite le potentiel performatif (Cabantous, Gond, 2011) de ces concepts, en analysant l'alignement entre les trois dimensions précédemment mentionnées.

Cette recherche met en évidence les difficultés d'articulation entre les différents concepts étudiés, notamment en raison de la concurrence entre les acteurs impliqués et les dispositifs d'action préconisés. Les résultats présentent une double contribution. D'une part, nous contribuons à la littérature existante sur les défis de l'adoption de l'IA et de son inscription dans une démarche d'innovation responsable, notamment au regard de la performativité des concepts d'IA éthique (Brundage, 2016 ; Stahl, Wright,

2018 ; Grinbaum, 2018 ; Buhmann, Fieseler, 2021, 2022 ; Stahl, 2022). D'autre part, nous élargissons cette réflexion au-delà de l'IA, en examinant les risques et les opportunités liés à la multiplication de concepts pour intégrer les technologies émergentes au développement multipolaire dans une démarche d'innovation responsable (Brundage, 2016 ; Stahl, 2022).

Revue de littérature

L'innovation responsable face au défi de l'intelligence artificielle

Face aux défis croissants suscités par des technologies émergentes, le concept d'innovation responsable (IR) – qui désigne une démarche visant à mieux articuler processus d'innovation et responsabilités sociales et sociétales – gagne en importance depuis une décennie (Ribeiro *et al.*, 2017). Alors que sa définition précise est sujette à débat (Ribeiro *et al.*, 2017 ; Stahl, Wright, 2018 ; Debref *et al.*, 2019), l'innovation responsable traduit consensuellement deux nécessités : gouvernance et anticipation. Von Schomberg (2013) souligne ainsi l'importance d'inscrire l'innovation dans un processus transparent et interactif par lequel les acteurs sociétaux et les innovateurs deviennent mutuellement réactifs les uns envers les autres en vue de l'acceptabilité (éthique), de la durabilité et de la désirabilité sociétale du processus d'innovation et de ses produits commercialisables, afin de permettre une bonne intégration des avancées scientifiques et technologiques dans notre société. Dans cette veine, l'Union européenne décrit ainsi l'innovation responsable comme « *la transition vers une nouvelle situation – et une amplification des possibilités – pour répondre aux obligations et honorer plus de devoirs envers les autres êtres humains, l'environnement, la planète et les générations futures qu'auparavant* » (Kormelink, 2019, p. 11). Pour cela, certains auteurs insistent sur l'importance de s'inscrire dans une démarche anticipatrice, à l'instar de Stilgoe *et al.* (2013, p. 1570) qui définit l'innovation responsable comme le fait de « *prendre soin de l'avenir par une gestion collective de la science et de l'innovation dans le présent* ». Cette approche de l'innovation responsable élargit la discussion sur la gouvernance pour englober les questions d'incertitude (sous ses multiples formes), les objectifs, les motivations, les trajectoires et directions de l'innovation ainsi que la dimension collective de la responsabilité de l'innovation. Le cadre proposé pour soulever,

discuter et répondre à ces questions se structure alors autour de quatre dimensions : anticipation, réflexivité, inclusion, réactivité (Stilgoe *et al.*, 2013 ; Ribeiro *et al.*, 2017).

Dans cette perspective, l'IA représente un défi pour le concept d'innovation responsable. D'une part, l'innovation responsable doit être capable d'influencer positivement le développement de l'IA (Brundage., 2016 ; Stahl, Wright, 2018 ; Buhmann, Fieseler, 2021, 2022). D'autre part, certains auteurs soulignent aussi la nécessité de faire évoluer certaines conceptualisations de l'innovation responsable pour adapter le concept aux spécificités de technologies comme l'IA (Grinbaum, 2018 ; Buhmann, Fieseler, 2021, 2022 ; Stahl, 2022).

Ainsi, l'intelligence artificielle, définie comme « *la capacité d'un système à interpréter correctement des données externes, à apprendre à partir de ces données et à utiliser ces apprentissages pour atteindre des objectifs et réaliser des tâches spécifiques grâce à une adaptation souple* » (Kaplan, Haenlein, 2019, p. 17), est une technologie (ou un ensemble de technologies liées) dont le potentiel de disruption est anticipé comme massif pour presque l'intégralité des strates socio-productives de notre société (Brynjolfsson, Mitchell, 2017 ; Makridakis, 2017 ; Kaplan, Haenlein, 2019). Au même titre que d'autres technologies émergentes avant elle (Te Kulve, Rip, 2011), l'ampleur des transformations induites par l'IA constitue une épreuve ayant valeur de test pour le concept d'innovation responsable et sa capacité à générer une démarche méliorative permettant d'orienter en amont les technologies émergentes au service du collectif (Brundage, 2016 ; Buhmann, Fieseler, 2021).

Or l'IA se caractérise par plusieurs spécificités qui questionnent la littérature en innovation responsable. L'une d'entre elles est liée à son développement polycentrique, impliquant des frontières technologiques, industrielles et géographiques complexes et mouvantes, qui oblige à repenser le concept d'écosystèmes (au pluriel) d'innovation responsables (Stahl, 2022, p. 31) : « *les écosystèmes d'IA peuvent être divisés par géographie (par exemple, européens, américains, chinois), ils peuvent être distingués par technologie (par exemple, apprentissage automatique, traitement du langage naturel, logique floue) ou par domaine d'application (par exemple, transport, santé, divertissement)* ».

Ce polycentrisme est d'ailleurs notamment l'une des raisons pour lesquelles les concepts éthiques sont aussi nombreux en IA. En réponse aux défis posés par l'intelligence artificielle, la littérature académique, souvent en lien avec le monde industriel, a en effet vu émerger un

bourgeonnement de concepts, tels que *Trustworthy AI*, *Sustainable AI*, *AI for Social Good*, ou encore *Fair AI*, pour n'en citer que quelques-uns (Mittelstadt *et al.*, 2016 ; Floridi *et al.*, 2018 ; Wamba *et al.*, 2021). Cette multiplication des concepts traduit la multiplicité des problématiques que soulève cette technologie, mais aussi la diversité des acteurs qu'elle bouscule, de leurs valeurs et des moyens qu'ils envisagent pour y remédier (Mittelstadt *et al.*, 2016 ; Floridi *et al.*, 2018 ; Adadi, Berrada, 2018 ; Stahl, 2021). Le risque de voir ces concepts rester de pures abstractions, des vœux pieux sans impacts réels sur le développement de cette technologie, préoccupe tout particulièrement le monde académique (Mittelstadt *et al.*, 2016 ; Edward, Veale, 2017 ; Mittelstadt, 2019 ; Martin, 2019 ; Owens, Walker, 2020 ; Santoni de Sio, Mecacci 2021 ; Merhi, 2022).

La multiplicité des concepts reflète à la fois la diversité des cas de figure et des parties prenantes de l'IA, mais peut aussi révéler une absence d'impact réel (Mittelstadt *et al.*, 2016 ; Brundage, 2016 ; Mittelstadt, 2019). Contrairement à la médecine, qui a réussi à concrétiser une bonne part de ses principes éthiques sur le terrain, dans le domaine de l'IA la traduction de ces concepts en actions concrètes reste incertaine : « *Le développement de l'IA ne dispose pas de méthodes empiriquement prouvées comparables pour traduire les principes en pratique dans des contextes de développement réels. Il s'agit d'un défi méthodologique aux multiples facettes* », souligne Mittelstadt (2019, p. 507). Certains auteurs mettent ainsi en garde contre le risque que l'IA ne parvienne pas à intégrer l'innovation responsable en multipliant des concepts qui n'ont pas d'impact sur la réalité des acteurs développant et commercialisant des IA (Brundage, 2016).

De la théorie à la pratique : les concepts d'IA responsable face au défi de leur performativité

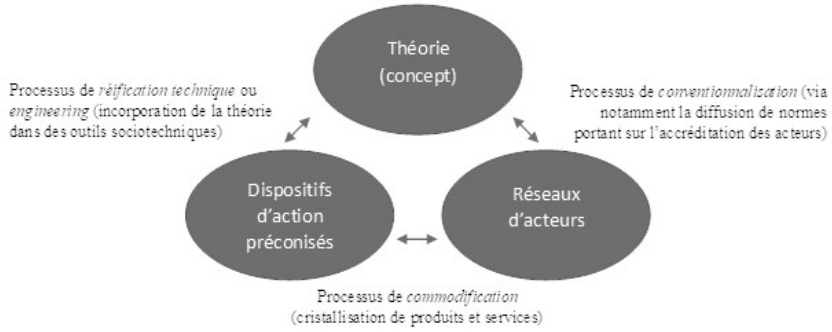
La multiplication de concepts reliés à l'éthique dans le domaine de l'IA offre une occasion favorable pour la promotion de l'innovation responsable, bien que cela représente également un défi théorique. Le problème de l'inscription de l'IA dans une démarche d'innovation responsable ne réside alors pas dans l'absence de concepts éthiques pour la guider, mais plutôt dans la capacité à opérationnaliser ces concepts et à les articuler de manière cohérente au sein des écosystèmes d'innovation responsables (Brundage, 2016 ; Buhmann, Fieseler, 2021).

Dans cette perspective, l'utilisation de la théorie de la performativité dans le domaine des sciences de gestion constitue une approche

analytique pertinente pour étudier la traduction d'une théorie en pratique (Cabantous, Gond, 2011 ; Marti, Gond, 2018). Cette approche met en évidence la création de réseaux d'acteurs ainsi que l'utilisation d'outils techniques qui permettent d'opérationnaliser la théorie et d'influer sur les décisions prises au sein des organisations.

La performativité d'un concept se produit lorsque s'enclenche une dynamique intra et inter-organisationnelle dans laquelle une théorie, les acteurs qui la soutiennent, ainsi que les dispositifs qu'ils créent, interagissent et se renforcent mutuellement (Cabantous, Gond, 2011). Cette boucle performative est illustrée dans la figure 1 ci-dessous.

Figure 1 – Représentation schématique inspirée de Cabantous et Gond (2011) d'une boucle performative articulant un concept, ses réseaux d'acteurs et ses dispositifs d'action associés



Lorsqu'une telle boucle performative s'active, cela implique des relations nouvelles entre la théorie, les acteurs et les dispositifs. Les dispositifs réifient la théorie, et inversement la théorie est crédibilisée par la présence d'outils dans ce sens (processus dit d'*engineering*) ; les acteurs encouragent l'émergence de dynamiques socio-professionnelles liées à la théorie (*conventionalizing*) concourant à légitimer à la fois la théorie et les acteurs qui en sont tributaires ; celles-ci participent à réguler, mais aussi structurer la production des dispositifs mettant en œuvre cette théorie et leur commercialisation (processus de *commodifying*). Lorsque la théorie, les acteurs et les dispositifs s'alignent, émerge alors une boucle performative qui favorise l'opérationnalisation du concept au sein des systèmes socioproductifs qu'il vise à transformer.

L'IR propose plusieurs modalités intégratives pour la mise en application des concepts au sein des écosystèmes d'innovation (Ribeiro *et al.*, 2017), notamment Stahl qui propose la conceptualisation d'espaces

« *constituées par des activités, des acteurs et des normes* » (Stahl 2013, p. 709). La théorie performative a alors pour intérêt d'aider à caractériser dans le détail l'articulation entre ces différentes dimensions, et l'effet d'entrainement qui s'enclenche lorsque ces trois dimensions s'alignent au sein d'une boucle performative (Cabantous, Gond, 2011 ; Aggeri, 2017).

Or, un tel alignement est loin d'être acquis dans le cas de l'IA. Ainsi, des études de terrain comme celle de Vakkuri *et al.* (2022) montrent que les développeurs sont sensibilisés à un certain nombre de concepts éthiques, mais qu'aucune des entreprises étudiées ne s'est véritablement outillée pour adresser ces enjeux (Vakkuri *et al.*, 2022). La traduction des nombreux concepts devant contribuer à inscrire l'IA dans une démarche d'innovation responsable en actes concrets nécessite en effet la génération d'un cycle performatif qui combine les concepts, les acteurs qui les portent, et les dispositifs pour les réifier et les diffuser. En l'absence de telles boucles, il y a un risque significatif de voir ces concepts comme « *toothless* », selon l'expression de Rességuier et Rodriguez (2020), c'est-à-dire sans mordant, sans impact sur la réalité pratique des acteurs de l'IA au quotidien.

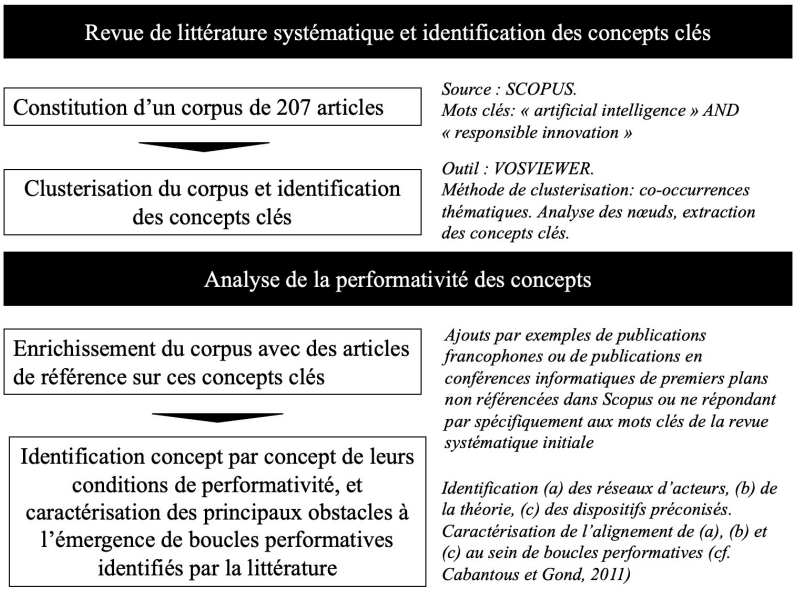
Face à la multiplication de concepts articulant IA et innovation responsable (Brundage, 2016 ; Buhmann, Fieseler, 2021, 2022 ; Stahl, 2022), un examen minutieux de leur capacité à opérer un impact dans les organisations et les marchés qui portent cette technologie est essentiel. Il est nécessaire, dans cet article, de dresser une liste de ces concepts, d'identifier les problématiques auxquelles ils tentent de répondre, de repérer les acteurs qui les soutiennent et les dispositifs qu'ils proposent de mobiliser à cette fin, et d'évaluer les principaux facteurs pouvant influencer l'émergence de ces boucles performatives de manière positive ou négative.

Méthode

Pour identifier les concepts clés et analyser leurs conditions de performativité, nous avons procédé en deux étapes. D'abord, une analyse bibliométrique de la littérature a permis d'identifier les concepts clés articulant IA et innovation responsable. Ensuite, nous avons analysé les conditions de performativité de ces concepts en caractérisant leurs boucles performatives (acteurs, théorie, dispositifs) et en identifiant les obstacles à leur émergence décrits dans la littérature. Les sous-sections suivantes et l'annexe 1 détaillent ces étapes, tandis que la section *Résultats* résume

l'analyse en listant les trois pôles des boucles performatives associées à chaque concept (fondement théorique, dispositifs préconisés, réseaux d'acteurs) et en mentionnant les écueils identifiés.

Figure 2 – Synthèse de la méthodologie utilisée



Constitution du corpus, *clusterisation* et identification des concepts clés alliant IA et innovation responsable

Un corpus de 207 articles portant sur l'IA et l'innovation responsable a été collecté sur la base de données Scopus. Les mots clés constitutifs de ce corpus (critères : inclus dans les *keywords*, l'*abstract* et/ou le titre) sont : innovation responsable et intelligence artificielle.

Ce corpus a été analysé à l'aide de la méthode des co-occurrences thématiques du logiciel Vosviewer (Van Eck, Waltman, 2010 ; Wong, 2018 ; Dabić *et al.*, 2021). La figure 3 en annexe résume les principaux nœuds et *clusters* thématiques qui en ressortent. L'analyse de ces *clusters* met en évidence trois sous-groupes. Le premier sous-groupe traite des prérequis éthiques lors du processus de conception, incluant les enjeux de régulation et de gouvernance. Le deuxième sous-groupe se concentre sur

les changements induits par l'IA dans les systèmes productifs, en particulier sur le plan industriel, et la question de la soutenabilité de ces transformations. Enfin, le troisième sous-groupe examine les conséquences de l'IA et la capacité de l'innovation responsable à influencer son développement dans le but de servir le bien commun. Dans ces trois sous-groupes, nous identifions les concepts clés qui articulent l'IA et l'innovation responsable. Il s'agit, pour le premier sous-groupe portant sur l'éthique pendant le processus de conception, des concepts d'IA digne de confiance (*Trustworthy AI*), d'IA transparente (*Tranparency AI*), et d'IA explicable (*Explainable AI*). Au sein du second sous-groupe portant sur la soutenabilité des processus productifs, deux concepts ont été identifiés : l'IA au service du développement durable (*AI for sustainability*), c'est-à-dire l'utilisation d'IA pour améliorer la durabilité de processus productifs ; et l'IA durable (*Sustainable AI ou Green AI*), c'est-à-dire la capacité à rendre l'IA moins consommatrice de ressources et d'énergie dans son fonctionnement. Enfin, le troisième sous-groupe questionnant la gestion des *outputs* de l'IA, et notamment le partage de la valeur, est porté en particulier par deux concepts clés : l'IA juste (*Fair AI*), et l'IA pour le bien commun (*AI for Social Good*).

Après avoir identifié les concepts clés grâce à la *clusterisation*, nous avons enrichi le corpus avec des travaux portant sur ces concepts. Cela inclut des articles provenant de conférences majeures en informatique qui ne sont pas répertoriées dans Scopus, ainsi que des articles en français qui n'étaient pas inclus dans notre corpus initial. Cette démarche d'enrichissement permet d'obtenir une revue de littérature plus complète pour chaque concept, ce qui facilite une analyse systématique des conditions de performativité.

Analyse des conditions de performativité par concept

La perspective performative invite à caractériser la capacité des concepts à devenir opérationnels. Pour cela, il s'agit d'identifier le fondement théorique de ces concepts, les acteurs qui les prônent, et les dispositifs qui appuient le concept (Cabantous, Gond, 2011) :

- La théorie : chaque concept propose une lecture particulière du problème. Les concepts résument ainsi souvent la difficile articulation entre IA et innovation responsable et mettent en avant un ou deux obstacles clés à résoudre. Nous cherchons ainsi à identifier : à *quoi* ce concept prétend-il répondre ? Quels sont ses principaux objectifs ?

- Les acteurs : nous identifions les réseaux d'acteurs qui portent ce concept. Pour cela, nous cherchons à répondre à la question : *qui*, suivant cette littérature, est supposé porter ce concept ? Qui est présenté comme un agent pouvant remédier au problème posé ?
- Les dispositifs : les concepts listés s'appuient sur un ensemble de moyens codifiés pour parvenir à répondre au problème qu'ils mettent en lumière. Ces dispositifs sont le plus souvent portés par les acteurs identifiés *supra*. Au sein de la littérature, pour chaque concept, nous cherchons donc à caractériser : *comment* le concept prétend inscrire l'IA dans une démarche d'innovation responsable ?

Nous examinons ensuite l'alignement entre ces trois dimensions, ce qui favorise l'émergence de boucles performatives (Cabantous, Gond, 2011 ; Aggeri, 2017). Ces boucles sont essentielles pour concrétiser les concepts dans les marchés et les organisations qu'ils cherchent à transformer (MacKenzie, Millo, 2003 ; MacKenzie *et al.*, 2006 ; Cabantous, Gond, 2011). Nous répertorions les principaux obstacles identifiés dans la littérature qui peuvent entraver ces alignements entre acteurs, théories et dispositifs recommandés, en précisant s'ils concernent principalement le processus de commodification (articulation entre acteurs et dispositifs), la conventionnalisation (articulation entre théorie et réseaux d'acteurs), ou d'incorporation technique (articulation entre théorie et dispositifs).

Résultats

Dans cette partie, nous répertorions les principaux concepts de la littérature sur l'IA responsable en les regroupant en trois pôles[1] : 1. les concepts liés au processus de conception, 2. les concepts axés sur la dimension écologique, et 3. les concepts concernant les usages et le partage de la valeur ajoutée de l'IA. Bien qu'il y ait des recoupements, ces catégories présentent des thématiques distinctes en termes d'objectifs, de réseaux d'acteurs et de moyens préconisés (Mittelstadt *et al.*, 2016 ; Wachter *et al.*, 2017). Pour chaque concept, nous identifions les parties prenantes, les problématiques de l'innovation responsable auxquelles ils sont censés répondre, et les dispositifs d'action recommandés pour les mettre en œuvre. Ensuite, nous répertorions les principaux risques liés à l'émergence d'une boucle performative entre ces trois dimensions.

1. Issus de l'étape de *clusterisation* décrite plus haut.

Le tableau ci-après résume ces résultats détaillés dans les sections suivantes.

Tableau 1 – Synthèse des principaux concepts de la littérature sur l'IA responsable

		Promoteurs principaux	**Problématiques adressées**	**Dispositifs promus**
IA et processus de conception	IA digne de confiance (*Trustworthy AI*)	Ethiciens, observatoires et comités éthiques (inter- et intra-organisationnels)	Répondre aux problèmes d'acceptabilité de l'IA, au niveau sociétal (peur de la machine autonome), organisationnel (attribution des responsabilités) et professionnel (acceptabilité au sein de collectifs professionnels aux normes éthiques élevés notamment)	Travail sur la gouvernance inter- et intra-organisationnel, souvent encadré par la création ou le renforcement de comités éthiques
		Groupes d'experts réunis à l'initiative des législateurs et des régulateurs		Élaboration de charte de principes cadrant la conception des outils, et promotion de cette charte auprès des utilisateurs finaux
				Lobbying auprès des régulateurs
	IA transparente (*Tranparency AI*)	Milieux associatifs, ONG et certains mouvements politiques souvent déjà investis sur les questions de discrimination ou liées aux données personnelles, sensibilisant à ces problématiques dans le cadre de l'IA les législateurs, les régulateurs, et certains comités éthiques	Discrimination envers certains groupes, souvent mais pas exclusivement minoritaires, en particulier sur des critères de genre, socio-ethniques, géographiques, d'opinion politique et d'orientations sexuelles	Renforcement des normes juridiques et des dispositifs de contrôle pour évaluer les biais, et ester en justice
		Organismes d'audit		Élaboration de méthodologie de contrôle pour organiser l'audit des outils d'IA et mettre en exergue des biais éventuels

		Promoteurs principaux	**Problématiques adressées**	**Dispositifs promus**
IA et processus de conception	IA explicable (*Explainable AI*)	Acteurs académiques et acteurs de l'industrie du développement des IA avec un tropisme technique	Répondre aux difficultés de compréhension (et donc de sécurité) et d'attribution des responsabilités liés à l'opacité de certains algorithmes alimentant aujourd'hui le boom de l'IA	Élaboration d'outils technique et d'interfaces associées permettant aux concepteurs et aux utilisateurs de mieux comprendre la machine (XAI)
		Éthiciens issus de domaines professionnels spécifiques comme la médecine		Standardisation de ces dispositifs et rédaction des normes afférentes
		Organismes de certification et de standardisation		
IA et soutenabilité des processus productifs	IA au service du développement durable (*AI for Sustainability*)	Acteurs académiques et associations professionnelles lançant un appel aux chercheurs en IA	L'IA est vue comme une opportunité pour répondre aux objectifs de développement durable, notamment pour réduire l'empreinte carbone et améliorer l'efficience de certaines activités	Chercheurs : production de cas d'usage, de schémas et données statistiques permettant de démontrer le potentiel de durabilité de l'IA. Production d'éléments rhétoriques appuyant les projets de R&D

		Promoteurs principaux	**Problématiques adressées**	**Dispositifs promus**
IA et soutenabilité des processus productifs	IA au service du développement durable (*AI for Sustainability*)	Acteurs économiques, particulièrement de secteurs polluants comme l'énergie, les transports, l'industrie, le BTP ; et les cabinets de conseils associés à ces practices	Des chercheurs font la promotion des usages vertueux de l'IA pour éclairer les milieux politiques. Les acteurs économiques voient l'utilisation de l'IA pour le développement durable comme une source de croissance économique et de conformité aux exigences écologiques, et les États comme un levier important de la « transformation verte et durable » dans une logique de croissance économique	Entreprises privées : cas d'usages, complétés par des estimations économiques et comptables
		Organismes nationaux et supranationaux issus de gouvernements, de l'UE, de l'ONU ou de l'OCDE, attachés aux problématiques de développement économique et de développement durable		États : hybridation entre discours politique et économique visant à une allocation des financements publiques et la création de dispositifs financiers et fiscaux en faveur de projets faisant levier sur l'IA pour le développement durable

		Promoteurs principaux	**Problématiques adressées**	**Dispositifs promus**
IA et soutenabilité des processus productifs	IA durable (*Sustainable AI*)	Acteurs académiques et associations professionnelles actifs dans le développement durable, l'économie circulaire, et les technologies « vertes »	L'idée centrale est que le développement et l'utilisation de l'IA doivent prendre en compte ses impacts écologiques sur l'ensemble de son cycle de vie « de manière à assurer une soutenabilité écologique forte de la planète ». L'objectif est de développer des outils d'IA plus frugaux que les systèmes qu'ils remplacent, et d'en rendre compte aux parties intéressées	Les dispositifs d'action se répartissent entre contraintes légales, édictions de principes à visée pratique sur le terrain, et des outils permettant de comptabiliser et rendre compte de l'impact économique, énergétique et environnemental de l'IA à différentes étapes de son cycle de vie

		Promoteurs principaux	**Problématiques adressées**	**Dispositifs promus**
IA et partage de la valeur	IA juste (*Fair AI*)	Laboratoires de recherche interdisciplinaires, attachés à des conférences (FATE notamment) et/ou opérant au sein de départements de R&D de grands acteurs technologiques (Microsoft, IBM, etc.)	Considérer les problèmes éthiques engendrés par l'IA, notamment les discriminations, avec une perspective plus systémique, prenant en compte les effets directs mais aussi indirects des algorithmes et leur inscription dans le temps dans le tissu sociétal et organisationnel	Promotion de la multidisciplinarité des approches
	IA pour le bien commun (*AI for social good*)	Acteurs économiques, associations professionnelles et syndicats (notamment patronaux)	Montrer que derrière les risques liés à l'IA, il y a aussi de formidables opportunités offertes par cette technologie pour la société, dans lesquels il faut investir rapidement	Promotion des initiatives vertueuses
		Agences nationales et internationales investies dans le développement économique	Orienter l'attention et les efforts humains et financiers vers les initiatives contribuant au bien commun et non seulement à des intérêts privés : infrastructure, éducation, etc.	Soutien à des mesures d'investissement financier et développement de mécanismes de financement *ad hoc*
				Orientation des financements vers des secteurs contribuant au collectif (éducation, infrastructure)

IA responsable et processus de conception

IA digne de confiance (*Trustworthy AI*)

Réseaux d'acteurs. L'IA digne de confiance, ou *Trustworthy AI*, est portée par des acteurs liés à des observatoires et comités éthiques, qu'ils soient indépendants ou affiliés à des organisations ou lobbys (Jobin *et al.*, 2019 ; Tiell, 2019 ; Hagendorff, 2020 ; Tang, 2020 ; Prunkl *et al.*, 2021). Certains de ces organismes sont créés par les législateurs, tels que le Groupe d'experts de haut niveau sur l'IA (IA HLEG) nommé par la Commission européenne. Initialement ancré en Europe (Floridi, 2019 ; Cohen *et al.*, 2020), le concept de l'IA digne de confiance est progressivement adopté aux États-Unis par le White House Office for Science and Technology (OSTP) et en Chine par les Beijing AI Principles (BAAI), ainsi qu'au niveau international par des instances comme l'OCDE (Rieder *et al.*, 2021).

Fondements théoriques et problématisation des enjeux. L'objectif de l'IA digne de confiance est double. (1) Le concept prend souvent en compte les préoccupations liées à l'acceptabilité de cette technologie (McClure, 2018), influencées par des œuvres fictionnelles emblématiques (Metropolis, Cycle des Robots, Matrix…) qui ont renforcé la méfiance de la société envers le risque de perte de contrôle face aux machines autonomes (Wosk, 2010). L'enjeu est donc de trouver des conditions qui réduiraient les réticences de la société envers le développement de cette technologie. (2) L'objectif corollaire est d'obtenir une meilleure compréhension des attentes de la population en termes de régulation. Le législateur, confronté à une opinion publique changeante et à une technologie émergente encore en évolution, doit jongler entre le renforcement des obligations légales concernant la conception et la commercialisation de l'IA, et le maintien d'un cadre juridique attractif pour les organisations développant des IA (Weaver, 2018 ; Veale, Borgesius, 2021 ; Floridi, 2021). Les comités éthiques jouent un rôle crucial en délimitant les grands principes susceptibles d'influencer les législateurs, les régulateurs et les instances étatiques dans l'élaboration de lois, ainsi que les politiques d'investissement (Veale, Borgesius, 2021 ; Floridi, 2021). À la demande de la Commission européenne, l'IA HLEG a ainsi produit deux documents importants intitulés « *Lignes directrices éthiques pour une IA digne de confiance* » et « *Recommandations en matière de politiques et d'investissements pour une IA digne de confiance* ».

Dispositifs préconisés. En termes de dispositifs, les partisans de l'IA digne de confiance mettent l'accent sur la gouvernance et l'établissement de grands principes. Des comités éthiques jouent généralement un rôle central dans cette démarche.

La gouvernance peut être mise en place au sein d'une organisation. Par exemple, Microsoft a créé un comité appelé « AETHER » (*AI, Ethics, and Effects in Engineering and Research*) ainsi qu'un département dédié à l'IA responsable (*Office of Responsible AI*). De même, Google a rendu publique l'initiative de son comité éthique, l'*Advanced Technology Review Council*, lorsque celui-ci s'est opposé à l'élargissement des palettes d'émotions identifiées par les IA de reconnaissance faciale en cours de recherche et développement. Le Conseil avait souligné que l'expression de la gêne ou du consentement était culturellement située, ce qui posait un risque de biais significatif. Dans la même optique, un comité éthique a été créé en 2019 à l'initiative de la ministre française des Armées Florence Parly, notamment pour aborder les problématiques liées à l'IA (Parly, 2019). Ce comité a conditionné l'utilisation et le développement de systèmes d'armes autonomes à un contrôle humain, en suivant le principe du « *Human in the Loop* » (Zanzotto, 2019 ; Grønsund, Aanestad, 2020). Ce principe est également promu par d'autres groupes de réflexion éthique qui soutiennent le concept d'IA digne de confiance, que ce soit dans d'autres armées, comme aux États-Unis (Gunning, 2017), ou dans le domaine médical (Holzinger *et al.*, 2022).

Les promoteurs de l'IA digne de confiance visent à étendre ce principe à toute une industrie. Les comités éthiques et les chartes qu'ils établissent agissent en tant que méta-organisations. Florence Parly, ministre des armées, souligne la nécessité d'une IA digne de confiance non seulement au sein de l'armée, mais également au sein de l'écosystème militaro-industriel : « *À ma demande, la DGA élabore actuellement un guide de développement maîtrisé des systèmes d'IA. Nous le partagerons et le consoliderons avec nos industriels, les laboratoires et toute la communauté des systèmes critiques, afin que nos armées puissent utiliser ces systèmes en confiance et en toute responsabilité* », insiste la Ministre (Parly, 2019).

Une critique adressée à de telles initiatives est le risque d'« *ethics washing* » et même d'« *ethics shopping* », où les comités éthiques méta-organisationnels imposeraient des contraintes peu problématiques pour les milieux professionnels concernés, utilisant ces efforts minimaux pour dissimuler les véritables enjeux (Wagner, 2018 ; Bietti, 2021).

IA transparente (*Transparency in AI*)

Réseaux d'acteurs. L'IA transparente est soutenue principalement par des associations, des ONG et certains mouvements politiques. Ces groupes, préexistants à la question de l'IA, prônent généralement la transparence dans d'autres domaines (lutte contre les discriminations, etc.) (Jobin *et al.*, 2019 ; Larsson, Heintz, 2020 ; Robinson, 2020). Le concept est également relayé de manière secondaire par deux autres groupes d'acteurs : les législateurs et les comités éthiques. Cependant, ceux-ci le considèrent davantage comme un modus operandi de leur propre programme plutôt que comme une finalité en soi (Jobin *et al.*, 2019 ; Larsson, Heintz, 2020 ; Robinson, 2020 ; Mora-Cantallops *et al.*, 2021).

Fondements théoriques et problématisation des enjeux. Le *Big Data* et l'IA font craindre l'émergence d'une société « boite noire » (Pasquale, 2015), où les individus seraient tributaires de décisions algorithmiques auxquelles ils ne pourraient ni se soustraire ni s'opposer (von Eschenbach, 2021). L'IA transparente vise alors à faciliter l'accès par toutes les parties prenantes aux algorithmes qui président à certaines décisions les affectant (Larsson, Heintz, 2020 ; Diakopoulos, 2020 ; Robinson, 2020 ; Walmsley, 2021). Cette exigence n'est pas spécifique à l'IA, et fait écho à une demande accrue de transparence au sein des organisations comme entre les organisations et leurs parties prenantes (Birchall, 2011 ; Hansen, Flyverbom, 2015). La transparence permettrait ainsi d'empêcher les arrangements secrets et répréhensibles en les soumettant au contrôle des individus, et d'autre part obliger les organisations à être plus rigoureuse en anticipation de ce potentiel contrôle (Felzmann *et al.*, 2020 ; Molina Rodríguez-Navas *et al.*, 2021).

Dans cette optique, les biais algorithmiques sont une préoccupation majeure liée à l'IA contre laquelle la transparence doit lutter. Cela inclut les biais de genre, ethniques, politiques, culturels, religieux, géographiques, socioéconomiques, etc. (Ntoutsi *et al.*, 2020 ; Daneshjou *et al.*, 2021). La transparence permet d'éviter des effets de discrimination trop prononcés et, le cas échéant, de les évaluer et de les dénoncer sur le plan politique ou juridique (Martin, 2019 ; Liu *et al.*, 2022 ; Bagaric *et al.*, 2022). L'importance du concept de *Transparency AI* se manifeste au sein des acteurs engagés dans les luttes contre les discriminations, ainsi qu'auprès des législateurs et régulateurs (Larsson, Heintz, 2020 ; Diakopoulos, 2020 ; Ntoutsi *et al.*, 2020 ; Robinson, 2020 ; Daneshjou *et al.*, 2021). Les comités éthiques intègrent fréquemment la transparence comme l'un des piliers de leurs recommandations afin de répondre aux exigences de transparence

de ces acteurs et de les anticiper. Selon Jobin *et al.* (2019), la transparence est le critère le plus présent parmi les principes des 89 comités éthiques examinés. Toutefois, contrairement aux ONG, mouvements associatifs et politiques prônant l'IA transparente, ces comités ne considèrent généralement pas la transparence comme une finalité, mais plutôt comme un moyen de faciliter l'IA digne de confiance (Jobin *et al.*, 2019 ; Larsson, Heintz, 2020 ; Mora-Cantallops *et al.*, 2021).

Dispositifs préconisés. Les partisans de l'IA transparente mettent en avant l'obligation légale, l'audit et la mise en place de dispositifs juridiques permettant de chiffrer et de contester les discriminations perçues (Castets-Renard, 2018 ; Bourcier, De Filippi, 2018 ; Martin, 2019 ; Robinson, 2020 ; Veale, Borgesius, 2021 ; Bagaric *et al.*, 2022). La capacité à procéduraliser en vertu du droit de transparence joue un rôle crucial dans le succès ou l'échec du concept d'IA transparente, car elle détermine si les groupes et associations qui en font la promotion peuvent se permettre les coûts juridiques associés (Washington, 2018). L'accès aux données est un enjeu clé (Bourcier, De Filippi, 2018 ; Veale, Borgesius, 2021), notamment en ce qui concerne la protection du secret industriel (Castets-Renard, 2018).

IA explicable (*Explainable AI*)

Réseaux d'acteurs. L'*Explainable AI*, ou XAI, réunit des acteurs académiques et de l'industrie du développement des IA (Gunning, 2017 ; Adadi, Berrada, 2018 ; Arrieta *et al.*, 2020). Ils proviennent de domaines variés où l'explicabilité de l'IA est une préoccupation majeure en matière de sécurité, tels que la robotique et l'armée (Gunning, 2017), ainsi que des éditeurs de logiciels tels que Microsoft Azure et IBM Watson. Les principaux acteurs du domaine sont initialement issus de l'informatique et des mathématiques (Adadi, Berrada, 2018 ; Arrieta *et al.*, 2020), mais récemment, des éthiciens, notamment issus de domaines professionnels comme la médecine (Holzinger *et al.*, 2022 ; Shaban-Nejad *et al.*, 2021 ; Antoniadi *et al.*, 2021), contribuent à nourrir et à orienter l'XAI vers une perspective davantage centrée sur l'utilisateur (Meske *et al.*, 2022 ; Zednik, 2021 ; Watson, Floridi, 2021).

Fondements théoriques et problématisation des enjeux. Le problème central de l'*XAI* est l'opacité de certains modèles d'IA qui résultent d'un grand nombre d'opérations mathématiques difficiles à retracer pour les humains comme dans le cas des réseaux de neurones artificiels (Goodfellow *et al.*, 2016 ; Gunning *et al.*, 2019). Cette opacité est

souvent négligée par les concepts précédents (Lipton, 2016). L'XAI est principalement axé sur des aspects techniques, développant des outils et des dispositifs (souvent algorithmiques) pour améliorer l'explicabilité des algorithmes d'IA (Gunning *et al.*, 2019 ; Arrieta *et al.*, 2020). Il succède au domaine académique de l'*IML* (*Interpretable Machine Learning*, Molnar *et al.*, 2020). Alors que l'IML se concentre sur la compréhension des mécanismes des algorithmes d'apprentissage automatique, par exemple pour le débogage ou l'amélioration de la fiabilité, l'XAI intègre cette réflexion tout en prenant en compte les autres parties prenantes externes, notamment les utilisateurs finaux, dont les attentes et les priorités peuvent différer considérablement de celles des développeurs informatiques (Arrieta *et al.*, 2020).

Dispositifs préconisés. L'XAI promeut deux principaux dispositifs. Le premier consiste en la création d'outils techniques pour améliorer l'interprétabilité des algorithmes après leur phase d'apprentissage, tels que des cartes de chaleur indiquant les parties de l'image utilisées par l'IA pour sa classification (Adadi, Berrada, 2018 ; Arrieta *et al.*, 2020 ; Wehbe *et al.*, 2021). Ces outils cherchent également à adapter l'explication en fonction des perspectives des parties prenantes (Watson, Floridi, 2021). L'XAI joue un rôle crucial dans la gestion de projets d'IA et la conception des structures décisionnelles, favorisant l'interaction entre les développeurs et les experts métiers (Shrestha *et al.*, 2019 ; Grønsund, Aanestad, 2020 ; Arrieta *et al.*, 2020).

Le deuxième dispositif émergeant de l'XAI est la production de normes industrielles. Cette standardisation est menée au sein d'agences et de consortiums internationaux, en collaboration avec l'industrie et sous l'impulsion de la Commission européenne, visant à établir des normes standardisées (ISO/IEC JTC1/SC 42 Artificial intelligence) (Zielke, 2020 ; Larsson, Heintz, 2020 ; Hagendorff, 2020). La législation en cours en Europe, notamment l'IA Act, renforce le rôle de cette standardisation de l'explicabilité des outils. Les producteurs et distributeurs d'IA ont le choix d'interpréter les exigences essentielles de la régulation en fonction du niveau de risque associé à leur IA, ou de se conformer aux recommandations des organismes de standardisation mandatés par la Commission pour établir des normes dans ce domaine. Cependant, le statut de droit privé de certains de ces organismes soulève des questions sur leur place dans les dispositifs juridiques tels que l'IA Act européen (Veale, Borgesius, 2021).

IA responsable et soutenabilité des processus productifs

IA au service du développement durable (*AI for Sustainability*)

Réseaux d'acteurs. L'IA au service du développement durable (*AI for sustainability*) est soutenue par des acteurs engagés dans cette trajectoire de développement, tels que des chercheurs, des associations, des entreprises privées (PWC, 2019 ; Rambach, 2022), ainsi que des organismes nationaux et supranationaux, dont certains sont liés à l'ONU ou à l'UE (Khareghani, 2020 ; Vinuesa *et al.*, 2020 ; Rolnick *et al.*, 2019 ; Fournier-Tombs, 2021 ; Gailhofer *et al.*, 2021).

Fondements théoriques et problématisation des enjeux. *AI for sustainability* vise à promouvoir le développement durable en utilisant l'IA pour équilibrer les aspects environnementaux, économiques et sociaux (Mensah, 2019). Les acteurs de ce concept soutiennent principalement que l'IA offre une capacité de calcul et de traitement de l'information prometteuse pour le développement durable. Les entreprises perçoivent également l'IA comme une opportunité de contribuer aux objectifs du développement durable, en réduisant leurs émissions de gaz à effet de serre et en se conformant aux réglementations en vigueur, tout en favorisant leur croissance économique (PWC, 2019). Philippe Rambach de Schneider Electric déclare que l'IA permet de réaliser des engagements en matière de décarbonation, de développement durable et de responsabilité sociale, tout en favorisant la transformation digitale et durable des activités (PWC, 2019).

Dispositifs préconisés. Les dispositifs de l'*AI for sustainability* visent à développer et promouvoir des systèmes d'IA réduisant l'impact environnemental des activités polluantes. L'utilisation combinée de l'IA, de la blockchain et de l'internet des objets permet de rationaliser la circulation des biens et des personnes, notamment dans les villes connectées, les chaînes d'approvisionnement et les hôpitaux (Singh *et al.*, 2020 ; Toorajipour *et al.*, 2021 ; Benzidia *et al.*, 2021).

La promotion de l'IA au service du développement durable repose sur la présentation de cas d'usage démontrant son potentiel. Rolnick *et al.* (2019) proposent plusieurs cas d'usage montrant comment l'IA peut réduire les émissions de gaz à effet de serre et aider les sociétés à s'adapter au changement climatique dans treize secteurs d'activité. L'évaluation des impacts négatifs et positifs de l'IA est essentielle, en privilégiant son

utilisation lorsque le bilan est positif. Les entreprises privées utilisent également des estimations économiques et comptables pour présenter les cas d'usage, incluant la productivité, les gains de parts de marché, la création d'emplois, la réduction des coûts, l'augmentation du volume d'affaires et la diminution des émissions de carbone (PWC, 2019).

IA durable (*Sustainable AI*)

Réseaux d'acteurs. Le concept d'IA durable *(Sustainability of* AI), aussi appelée IA verte *(Green AI)*, est principalement porté par des chercheurs (Rolnick *et al.*, 2019 ; Schwartz *et al.*, 2020 ; Lannelongue *et al.*, 2021 ; van Wynsberghe, 2021 ; Wu *et al.*, 2022) ou des associations professionnelles (Schwartz *et al.*, 2020) soucieuses de l'impact écologique de l'IA.

Fondements théoriques et problématisation des enjeux. L'objectif de l'IA durable est de réduire l'impact environnemental de cette technologie. Aujourd'hui les plus grands modèles d'IA sont à l'origine de la consommation d'une grande quantité d'énergie et de ressources pour collecter et stocker d'immenses volumes de données, ainsi que pour effectuer des calculs intensifs pendant les phases d'apprentissage (Murdock, Brevini, 2019 ; Brevini, 2020 ; Dauverge, 2020). L'objectif n'est donc plus seulement d'utiliser l'IA pour le développement durable, comme le propose le concept d'*AI for Sustainability*, mais de rendre l'IA elle-même durable en réduisant la quantité de données nécessaires, les calculs et la consommation énergétique des infrastructures.

Dans cette optique, la Déclaration de Montréal pour un développement responsable de l'IA propose plusieurs mesures (Dilhac *et al.*, 2018). Elle recommande de viser une plus grande efficacité énergétique et de réduire les émissions de gaz à effet de serre, de minimiser les déchets électriques et électroniques et de prévoir des filières de maintenance, de réparation et de recyclage dans une perspective d'économie circulaire. Elle préconise également de limiter les impacts sur les écosystèmes et la biodiversité, en particulier lors de l'extraction des ressources naturelles et en fin de vie. Enfin, elle appelle les acteurs publics et privés à soutenir le développement de systèmes d'IA écologiquement responsables afin de lutter contre le gaspillage des ressources naturelles, d'établir des chaînes d'approvisionnement durables, de favoriser les échanges commerciaux soutenables et de réduire la pollution à l'échelle mondiale.

Dispositifs préconisés. Les outils de l'IA durable comprennent des contraintes légales, des principes pratiques sur le plan juridique et des outils de comptabilisation pour évaluer l'impact économique, énergétique

et environnemental de l'IA à différentes étapes de son cycle de vie (Linkov *et al.*, 2018). D'abord, le principe d'« éco-conditionnalité », introduit par le Premier ministre français le 9 juillet 2013, conditionne l'octroi d'aides publiques ou d'avantages fiscaux au respect de critères environnementaux. Dans le contexte des projets d'IA, cela signifie qu'il faut placer la maîtrise du rapport coût/bénéfice environnemental au cœur de la conception du projet. D'autre part, des travaux de recherche proposent différentes méthodes pour mettre en pratique ces principes, telles que des outils de calcul pour mesurer la consommation énergétique et l'empreinte carbone, des cadres méthodologiques pour les soutenir, ainsi que des recommandations pour réduire l'empreinte carbone. Par exemple, Schwartz *et al.* (2020) proposent de calculer le nombre d'opérations de base nécessaires pour obtenir un résultat informatique (*floating point operations – FPO*). Henderson *et al.* (2020) proposent un autre outil de suivi de l'impact environnemental des expériences (*Experiment impact tracker*), qui intègre plusieurs variables liées à la consommation énergétique des composants informatiques, à l'efficacité énergétique du centre de données et à sa localisation.

IA responsable et partage de la valeur

IA juste (*Fair AI*)

Réseaux d'acteurs. Le concept d'IA juste est porté par des groupes multidisciplinaires tels que la conférence FATE (*Fairness, Accountability, Transparency, and Ethics in AI*), issue de la Web Conference et de l'ACM (*Association for Computing Machinery*) (Adadi, Berrada, 2018 ; Bird *et al.*, 2020 ; Mehrabi *et al.*, 2021). Au sein de grandes entreprises informatiques comme Microsoft ou IBM, des groupes de travail réunissant chercheurs et praticiens se revendiquent également affiliés à la conférence FATE (Bellamy *et al.*, 2019 ; Bird *et al.*, 2020).

Fondements théoriques et problématisation des enjeux. Le concept propose d'aborder les problématiques éthiques engendrées par l'utilisation d'IA avec une focale relativement large et interdisciplinaire, afin de capturer des effets, notamment de discriminations, qui seraient systémiques et donc non directement observables à l'échelle d'un individu. L'objectif est alors de proposer des pistes pour anticiper et amoindrir les conséquences indésirables liées à l'utilisation d'IA, avec une focale spatiale et temporelle étendue. Ici, la focale diffère de celle mobilisée par les concepts précédents. Dans l'IA digne de confiance, de transparence et d'explicabilité, la

focale porte surtout sur les dimensions directement mesurables à l'échelle d'un algorithme spécifique ou sur un groupe d'individu dans une situation donnée – comme par exemple la détermination statistique d'une discrimination sur des critères de profilage socio-ethnique, par exemple dans le cas de l'outil prédictif COMPASS utilisé par la justice américaine pour les remises de peine (Washington, 2018 ; Martin, 2019). A contrario, le concept de *Fairness* considère les effets systémiques de la discrimination sur les populations, qui peuvent être causés non seulement par l'IA elle-même, mais également amplifiés par celle-ci. Il tient compte de l'intersectionnalité, c'est-à-dire des facteurs multiples et interconnectés qui contribuent à la discrimination (Feuerriegel *et al.*, 2020 ; Mehrabi *et al.*, 2021 ; Madaio *et al.*, 2022).

Dispositifs préconisés. Le concept d'IA juste valorise la multidisciplinarité, offrant une approche holistique des effets de l'IA par rapport aux concepts d'IA digne de confiance, transparente ou explicable, jugés utiles mais segmentés (Feuerriegel *et al.*, 2020 ; Wachter *et al.*, 2021 ; Mehrabi *et al.*, 2021 ; John-Mathews *et al.*, 2022 ; Madaio *et al.*, 2022). L'AI *Fairness* considère la justice comme un idéal plutôt qu'un objectif opérationnel à court terme (Dignum, 2021 ; Wachter *et al.*, 2021), soulignant que son objectif principal est d'éviter l'aggravation de situations problématiques existantes (Bennett, Keyes, 2020).

Les partisans de l'IA *Fairness* développent des toolkits et des checklists qui ont une vocation méliorative plutôt que normative. Ces outils favorisent la prise en charge de l'idéal de justice par les parties prenantes, en fournissant une infrastructure organisationnelle pour formaliser les processus ad hoc et autonomiser les défenseurs individuels (Bellamy *et al.*, 2019 ; Madaio *et al.*, 2022). Ils contribuent également à redéfinir le concept de justice en collaboration avec les parties prenantes plutôt qu'institutionnellement (John-Mathews *et al.*, 2022). Dans une optique d'empowerment, la culture organisationnelle est présentée comme un levier critique pour intégrer le concept de justice dans les activités de conception et de déploiement des IA (Madaio *et al.*, 2022 ; Robert *et al.*, 2020 ; Landers, Behrend, 2022).

IA pour le bien commun (*AI for Social Good*)

Réseaux d'acteurs. L'*AI for social good* et les concepts qui y sont associés (*Tech for Good* et *AI for Common Good*) sont promus par des réseaux d'acteurs engagés à mettre en évidence les opportunités offertes par l'IA pour la société, et pas seulement les risques qui en résulteraient. Cela

inclut des acteurs économiques impliqués dans des initiatives sociétales, tels que des syndicats, des associations patronales, des associations professionnelles, ainsi que des entités nationales et supranationales comme des organes de l'ONU, de l'OCDE ou de l'UE (Taddeo, Floridi, 2018 ; Berendt, 2019 ; Cowls *et al.*, 2021 ; Bondi *et al.*, 2021 ; Umbrello, Van de Poel, 2021 ; Wamba *et al.*, 2021 ; Foffano *et al.*, 2022).

Fondements théoriques et problématisation des enjeux. Le mouvement *AI for Social Good* prend en compte, en plus des concepts liés à l'écologie, les effets macroéconomiques des technologies, notamment les inégalités Nord/Sud, pauvres/riches et entre différentes générations ou populations ayant un accès différencié aux NTIC (Taddeo, Floridi, 2018 ; Floridi *et al.*, 2020 ; Cowls, 2021 ; Cowls *et al.*, 2021 ; Wamba *et al.*, 2021 ; Foffano *et al.*, 2022). Contrairement à la plupart des concepts précédents, l'AI for social good ne se concentre pas uniquement sur les externalités négatives de l'IA, mais cherche plutôt à inverser la perspective éthique de l'IA en mettant en évidence les opportunités offertes par cette technologie pour résoudre ces grands problèmes sociaux et encourager des projets d'IA vertueux.

Dispositifs préconisés. Un des mécanismes encouragés par les tenants de l'*AI for Social Good* est la combinaison de la promotion institutionnelle et du soutien financier (Bondi *et al.*, 2021 ; Cowls *et al.*, 2021 ; Foffano *et al.*, 2022). Le concept d'*AI for social good* vise à être volontairement non contraignant, ou du moins le moins contraignant possible (Taddeo, Floridi, 2018 ; Hermann, 2021 ; Floridi *et al.*, 2020 ; Cowls *et al.*, 2021). En effet, ce concept est alimenté par l'enthousiasme pro-business/technologique qui caractérise une part importante de ses promoteurs (Umbrello, Van de Poel, 2021 ; Cowls *et al.*, 2021 ; Foffano *et al.*, 2022). Il repose sur l'idée que la technologie peut être un moteur de changement social collectivement bénéfique, sous certaines conditions infrastructurelles (Cowls, 2021 ; Bondi *et al.*, 2021 ; Foffano *et al.*, 2022).

Le concept d'*AI for Social Good* se réfère notamment à la notion de « commun » (que l'on retrouve dans une variante du concept : l'*AI for Common Good*) notamment popularisée par les travaux de la prix Nobel Elinor Ostrom (1990) et met en valeur la logique des communautés, capables de générer des externalités positives grâce à la technologie et de limiter les externalités négatives (Berendt, 2019 ; Bondi *et al.*, 2021).

Cependant, la force du concept de « commun », notamment sa capacité à être pensé à différentes échelles et dans différents contextes socio-économiques et géographiques, constitue également sa principale limite

(Bondi *et al.*, 2021 ; Holzmeyer, 2021 ; Cowls, 2021 ; Cowls *et al.*, 2021). Certains auteurs soulignent ainsi le risque que l'influence socio-économique partielle de certains acteurs très actifs dans la promotion de l'*AI for social good* imprègne le concept de leurs valeurs et de leurs attentes en termes d'évolutions sociales. En maintenant une certaine ambiguïté autour du concept de « commun », l'*AI for social good* risque de dissimuler certaines questions éthiques fondamentales, telles que le bien « pour qui ? » et en investissant les biens « de qui » ? (Cowls, 2021).

Enjeux de performativité des concepts d'IA responsables

L'étude des concepts d'IA responsables met en avant leur fort potentiel pour favoriser une innovation responsable, mais aussi les limites et les écueils auxquels ils peuvent être confrontés. La littérature met en évidence un alignement théorique : chaque concept est accompagné d'un dispositif qui légitime d'une manière ou d'une autre l'action des acteurs. Cependant, plusieurs risques susceptibles de compromettre l'émergence et la stabilité de ces boucles performatives sont identifiables. Nous contribuons à mieux comprendre ces risques et aux liens qu'ils entretiennent avec les processus de *commodification*, la *conventionnalisation*, ou d'*incorporation technique*. Nous détaillons cette contribution ci-dessous.

Un premier risque concerne l'opérationnalisation technique de certains concepts et la maturité des dispositifs recommandés pour les mettre en pratique. Le processus d'*engineering* est ainsi remis en question. Par exemple, l'idéal d'intelligibilité défendu par le concept d'IA explicable peut se heurter aux difficultés techniques réelles pour expliquer véritablement les algorithmes (Edward, Veale, 2017 ; Arrieta *et al.*, 2020). L'XAI rencontre ainsi des défis techniques pour proposer des outils à la hauteur des attentes liées à l'IA explicable, au point que certains auteurs préconisent d'envisager une approche d'innovation responsable qui assume le caractère « boîte noire » de l'IA à court terme (Grinbaum, 2018). Et même lorsque des outils d'explication existent, leur déploiement nécessite la formation d'un grand nombre d'acteurs, y compris de novices en matière d'IA, voire l'émergence de nouveaux métiers à l'interface entre la conception des IA et les utilisateurs finaux (Kellogg *et al.*, 2020 ; Holzinger *et al.*, 2022 ; Arrieta *et al.*, 2020). L'XAI représente ainsi un défi non seulement technique, mais aussi sociotechnique, car il implique la création d'outils capables d'expliquer adéquatement la machine, ainsi que le développement d'outils ou de protocoles organisationnels favorisant un dialogue entre les nombreuses parties prenantes (Meske *et al.*, 2022).

Un deuxième risque découle du défi de définir des normes à la fois suffisamment strictes et claires pour être applicables, tout en restant suffisamment flexibles pour ne pas entraver le développement de l'IA dans la pratique. La Commission européenne, dans la première version de l'IA Act publiée le 21 avril 2021, reconnaît la complexité pour le législateur d'établir une approche réglementaire équilibrée et proportionnée. Par exemple, l'IA transparente repose sur des normes de transparence qui s'appliqueraient aux organisations, mais l'accès aux données et leur intelligibilité une fois transmises posent des problèmes, notamment en raison du secret industriel (Bourcier, De Filippi, 2018 ; Larsson, Heintz, 2020 ; Veale, Borgesius, 2021). D'un autre côté, la faible normativité de certains concepts limite leurs potentialités (Cowls *et al.*, 2021 ; Van Nood, Yeomans, 2021). La malléabilité de concepts tels que l'*IA for Social Good* ou la *Sustainable AI* facilite leur adoption par les acteurs, mais peut également réduire leur impact sur la transformation sociale réelle (Murdock, Brevini, 2019 ; Cowls *et al.*, 2021 ; Holzmeyer, 2021). Parallèlement, l'incertitude quant au caractère normatif de ces concepts remet en question le rôle des acteurs chargés de les faire respecter, en l'absence d'une définition claire et définitive du statut de ces normes (Veale, Borgesius, 2021). L'ambiguïté du positionnement de ces acteurs chargés d'assurer la transparence est largement débattue dans la littérature (Jobin *et al.*, 2019 ; Felzmann *et al.*, 2019, 2020 ; Larsson, Heintz, 2020 ; Veale, Borgesius, 2021). Après la publication de la première version de l'IA Act en 2021, de nombreuses interrogations sont apparues quant au rôle des *notified bodies* chargés de contrôler les mesures de transparence (Veale, Borgesius, 2021 ; Lilkov, 2021 ; Mökander *et al.*, 2022).

Enfin, un troisième risque s'observe dans l'enrayement de boucles performatives au niveau du processus de « *commodifying* », c'est-à-dire de formalisation de biens et services marchands et monétisables à même de concourir à la diffusion du concept et des dispositifs associés, et de contribuer à l'attractivité des métiers qui en découlent (Cabantous, Gond, 2011). Le défi peut porter sur l'absence de dynamique en faveur de la création de ce marché… ou à l'inverse, sur la trop rapide structuration d'un marché et le rapide positionnement d'acteurs qui contreviennent au concept éthique préalablement formulé. Ainsi, la difficulté ne repose pas tant sur l'absence d'acteurs incités à promouvoir le concept, que sur la nature de leur incitation, et leurs propensions, une fois leurs initiatives agrégées, à distordre le concept initial. Un risque concomitant est de voir des acteurs défendre un concept parce que le déploiement des solutions préconisées pourrait leur permettre d'alimenter, en sous-main, leurs bases de données, et ainsi

de s'ouvrir de nouveaux marchés. Ainsi, à partir d'une étude sur les initiatives siglées AI4SG (*AI for Social Good*) dans le domaine de la santé, Holzmeyer (2021, p. 94) souligne que « *de nombreuses initiatives AI4SG sont portées par les mêmes entreprises qui incubent les technologies de l'IA […] encapsulent souvent les problèmes sociaux et environnementaux systémiques (…)* ».

Discussion

Performativité et innovation responsable : une grille analytique pour transformer la théorie en pratique

Notre étude dresse un constat en demi-teinte. Il souligne d'une part le fort potentiel des principaux concepts listés à faire émerger des boucles performatives afin de traduire la théorie en dispositifs concrets. En revanche, nos résultats mettent en garde envers certains risques pouvant, à moyen terme, enrayer ces boucles.

Dans cette optique, une première contribution vise la typologisation de ces difficultés, notamment en les décomposant selon qu'ils portent sur des processus dits d'*engineering*, de *conventionalizing* ou de *commodification*. Cette spécification permet de mieux articuler IA et innovation responsable (IR), en précisant les processus pouvant aider à la définition de solutions permettant d'éviter ou de surmonter ces écueils. Par exemple, l'innovation responsable a développé un ensemble conséquent de méthodes de délibération collective (Buhmann, Fieseler, 2021, 2022). Ces méthodes peuvent être utiles pour surmonter les problématiques de convergence d'audiences rencontrées dans le processus de *conventionalizing*, notamment les difficultés des promoteurs de l'IA explicable pour concevoir des protocoles organisationnels permettant de faire dialoguer des acteurs aux discours et aux attentes très différentes sur l'IA (Arrieta *et al.*, 2020).

Notre étude montre aussi l'intérêt de la théorie de la performativité comme grille analytique pour diagnostiquer l'opérationnalisation de concepts soutenant l'IR notamment au regard de sa modularité en termes d'échelles spatiotemporelles d'analyse. En effet, que ce soit pour l'IA (Brundage, 2016) ou des technologies dérivées comme les véhicules autonomes (Cohen *et al.*, 2018), la littérature en IR insiste de manière croissante sur la nécessité de trouver une ligne médiane entre d'une part un niveau de lecture trop abstrait, qui éluderait la problématique de la mise en œuvre pratique de l'innovation responsable au niveau des acteurs,

et d'autre part une vision trop locale sur les dispositifs d'action qui occulterait la dimension systémique de ces technologies et des risques qu'elles engendrent (Galaz *et al.*, 2021). De ce point de vue, en mettant l'accent sur l'alignement entre théorie, réseaux d'acteurs et dispositifs, ce travail offre alors une grille analytique permettant une transversalité d'échelle, englobant des blocages locaux et une vision plus systémique de l'IR. L'autre avantage est de pouvoir étudier l'opérationnalisation des concepts d'IR en se focalisant par exemple sur une région. Dans cette perspective, de plus en plus d'auteurs soulignent la nécessité de penser la multiplicité et les divergences en matière d'IR entre différents écosystèmes d'innovation pour une même technologie (Stahl, 2022). Notre étude suggère alors que la modularité de la grille d'analyse performative, et sa capacité à fournir un diagnostic situé et adapté à l'échelle d'analyse souhaitée, est un vrai atout pour conserver une unicité de méthode et une comparabilité des diagnostics entre segments d'analyse.

L'innovation responsable face à une multiplication des concepts éthiques

Une deuxième contribution consiste à examiner comment l'IR peut influencer le développement d'innovations émergentes dans un contexte polycentrique, où différents acteurs sont impliqués avec des rythmes et des ambitions variées, voire opposées. Une partie des travaux sur l'IR suppose que le développement technologique est effectué par un nombre limité d'acteurs, coordonnés ou non, au sein d'écosystèmes d'innovation identifiables (Brundage, 2016 ; Stahl, 2022). Même si d'autres technologies ont aussi été identifiées comme polycentriques, le niveau d'éclatement du processus de développement de l'IA oblige clairement à revoir la théorie pour rendre compte d'une IR impulsée à plusieurs niveaux, à plusieurs rythmes, par et au sein de multiples écosystèmes d'innovation (Stahl, 2022). Or, un tel niveau de dispersion des initiatives pour développer ces technologies induit aussi une dispersion des initiatives pour les améliorer. C'est particulièrement le cas de l'IA : alors que la littérature soulignait la multiplication des concepts pour rendre l'IA plus éthique (Mittelstadt *et al.*, 2016 ; Mittelstadt, 2019 ; Santoni de Sio, Mecacci, 2021), notre étude montre aussi que ces concepts sont appuyés par des réseaux d'acteurs divers, et promeuvent des dispositifs différents. Or, cette dispersion questionne : elle multiplie les regards, donc est plus à même de respecter la diversité des parties prenantes, mais limite aussi le champ d'action, alors que l'innovation responsable appelle initialement une portée plus globale de l'action. Comme le regrette Brundage (2016, p. 545), « *On a beaucoup*

écrit sur les dimensions sociétales de l'IA, mais ces ouvrages ont tendance à se concentrer sur des sous-thèmes discrets de questions sociales soulevées par l'IA, un à la fois, et à s'orienter vers des risques ou des opportunités découlant de l'IA, plutôt que sur la nécessité d'une approche systémique pour renforcer les capacités dans l'ensemble de l'écosystème émergent de la science et de l'innovation ».

Nos résultats reformulent le questionnement soulevé notamment par Brundage (2016) et Stahl (2022) sur la multiplication des initiatives pour une IA plus éthique en mettant en lumière le défi d'articuler efficacement les boucles performatives des concepts. Dans une perspective d'innovation responsable, la question qui émerge est : faut-il aider le renforcement de chaque concept séparément, en facilitant l'émergence de boucles performatives associées, au risque de les voir s'autonomiser, ou au contraire accélérer l'intégration des boucles performatives entre elles ?

En effet, si on applique la littérature sur la performativité (MacKenzie, Millo, 2003 ; MacKenzie *et al.*, 2007 ; Cabantous, Gond, 2011), l'émergence de boucles performatives indépendantes des autres concepts devraient le plus probablement avoir tendance à renforcer la triade théorie/acteurs/dispositifs. Les réseaux d'acteurs en charge du déploiement auraient intérêt à circonscrire leur juridiction de manière à garder le contrôle sur le déploiement des dispositifs préconisés. Autrement dit : ces concepts seraient en quelque sorte autonomes, générant un marché et des actions relativement indépendamment des autres concepts. Le risque est alors double : d'une part, de voir s'accentuer le constat de Brundage (2016) sur un morcellement de la question éthique de l'IA qui échoue à infléchir durablement et globalement le développement de cette technologie. Dans cette configuration, la multiplication des concepts et des initiatives éthiques serait tendanciellement contraire à la démarche d'innovation responsable. Le second risque est de voir émerger une forme d'*ethic shopping*. L'*ethic shopping* pour l'IA désigne une forme de cynisme institutionnalisé qui pousserait les acteurs à s'engager envers les concepts éthiques qui leur seraient les moins contraignants tout en délaissant les autres (Wagner, 2018 ; Bietti, 2021). Selon cette perspective, par exemple, les concepteurs d'IA pour la justice prédictive (Brayne, Christin, 2021), dont l'un des enjeux est d'assurer l'absence de biais (notamment envers les origines socio-ethniques supposés, Martin, 2019) préféreraient s'engager envers des concepts d'IA durable, dont les enjeux sont complètement orthogonaux aux réelles difficultés éthiques générées par cette technologie. Inversement, des concepteurs d'IA les plus polluantes s'engageraient

inversement en faveur d'IA explicable et transparente, etc. Cowls *et al.* (2021) et Holzmeyer (2021) mettent ainsi en garde contre une forme de parasitage de certains concepts par des acteurs aux motivations éthiques douteuses, qui circonscrivent le concept pour mieux l'opérationnaliser à leur fin.

À l'inverse, intégrer les concepts entre eux fait aussi courir d'autres risques. A priori, la multiplication des concepts peut apparaitre comme une bonne chose du point de vue de l'IR, car elle permet d'appréhender les multiples facettes de la technologie, ce qui est plus difficile avec un seul concept centralisé forcément réducteur (Brundage, 2016). Dans le cas de l'IA, un certain nombre de concepts semble se compléter. Par exemple, l'IA digne de confiance, l'IA transparente et l'IA explicable tendent à se renforcer mutuellement, de telle sorte que ces trois concepts se retrouvent souvent mentionnés ensemble, bien que priorisés différemment en fonction des acteurs qui les promeuvent (Floridi *et al.*, 2018 ; Jobin *et al.*, 2019 ; Hagendorff, 2020). En particulier, le développement de l'IA explicable est cité comme pouvant fortement contribuer à rendre l'IA transparente plus facilement opérationnelle, en lui fournissant des outils de contrôle de la machine (Arrieta *et al.*, 2020). De manière similaire, l'IA explicable et l'IA transparente sont listées comme des facteurs clés pour l'émergence d'une IA digne de confiance (Floridi *et al.*, 2018). De même, les concepts d'IA durable et d'IA au service du développement durable s'inscrivent aussi à l'agenda de l'*AI for Social Good* (Wamba *et al.*, 2021). Mais en pratique, il existe une forme de compétition entre ces concepts. Si l'IA durable, l'IA au service du développement durable, et l'IA pour le bien commun semblent porter des objectifs connexes, dans les faits, leurs préconisations en termes d'orientation des investissements (privés, gouvernementaux) et d'élaboration de dispositifs fiscaux éventuellement associés traduisent des priorités économiques qui les mettent en compétition (Cowls, 2021 ; Foffano *et al.*, 2022).

Ces concepts convergent donc, mais jusqu'à un certain point ; et l'on peut même imaginer un découplage à moyen terme. En privilégiant certains leviers et certains dispositifs d'action, les pouvoirs publics, les législateurs et les régulateurs, pourraient faire pencher la balance et modifier les équilibres entre ces concepts.

Limites et pistes pour de futures recherches

Notre étude contribue à une meilleure articulation entre IA et innovation responsable par l'apport de la théorie de la performativité. Elle intègre aussi certaines limites qui appellent de futures recherches.

Premièrement, méthodologiquement, notre étude se concentre sur les concepts les plus saillants à ce jour : or, la littérature, et notamment celles sur les concepts éthiques en IA, est extrêmement vivante (Mittestadt *et al.*, 2016 ; Mittelstadt, 2019 ; Santoni de Sio, Mecacci, 2021). Par exemple, le concept d'*accountability* de l'IA prend progressivement une importance grandissante dans la littérature, bien qu'encore subordonnée aux concepts listés dans ce travail. Alors que l'IA se développe, et commence à être intégré dans les organisations, de nouvelles difficultés sont découvertes. Pour certains auteurs, il s'agit même d'anticiper ses problèmes pour les circonscrire en amont (Munoko *et al.*, 2020). Il faut ajouter que la multiplication des écosystèmes d'innovation de l'IA, avec des frontières géographiques, techniques et industrielles instables (Stahl, 2022), oblige à reconnaitre le caractère continuellement évolutif de notre cartographie. Dans cette optique, si les résultats et les contributions pour la littérature de l'innovation responsable obtenue à ce jour restent méthodologiquement valides, l'image globale peut évoluer, avec l'apparition de nouveaux concepts appelant à un travail de réactualisation régulier – une problématique courante pour l'innovation responsable, mais particulièrement vive pour l'IA.

Parallèlement, les critères retenus pour notre étude orientent vers une lecture plutôt « occidentale » et centrée sur les industries les plus en pointe sur la question éthique de l'IA. Plusieurs travaux soulignent en effet l'orientation assez marquée des valeurs des concepts éthiques appliqués à l'IA (Ménissier, 2020 ; Adams, 2021). De même, l'influence notamment de la médecine (Beckers *et al.*, 2021 ; Holzinger *et al.*, 2022 ; Shaban-Nejad *et al.*, 2021 ; Antoniadi *et al.*, 2021) et d'autres industries comme l'audit (Munoko *et al.*, 2020) sur les questions éthiques de l'IA pourraient avoir tendance à produire un effet grossissant, occultant d'autres problématiques encore sous-jacentes à ce jour (Mittelstadt, 2019 ; Holzmeyer, 2021).

Notre étude aboutit à une nouvelle problématique : pour l'IA, l'enjeu de l'innovation responsable est d'équilibrer les concepts en évitant une autonomisation excessive et une intégration trop forte. Le défi est de créer la « méta-performativité » des concepts, c'est-à-dire de trouver

comment chaque concept pourrait contribuer aux boucles performatives des autres concepts. Notre étude est un premier pas vers cela, nécessitant des travaux ultérieurs. Les résultats soulignent l'importance de la théorie de la performativité pour l'innovation responsable, mais se limitent à identifier l'importance de l'étude de la méta-performativité sans répondre à la manière d'y parvenir.

Conclusion

Cet article propose une cartographie des concepts clés de l'IA pour l'innovation responsable, permettant de mieux comprendre leur place pour le monde académique et professionnel. En utilisant une approche analytique basée sur la performativité, nous examinons leur traduction en pratiques concrètes.

L'IA interroge l'influence de l'innovation responsable sur les innovations émergentes dans un contexte polycentrique, avec des acteurs aux ambitions diverses. Le défi réside dans l'articulation efficace des boucles performatives des concepts de l'IA. Dans cette perspective, la question qui émerge est : faut-il renforcer chaque concept séparément, en favorisant l'émergence de boucles performatives autonomes, ou accélérer leur intégration ? Les deux scénarios comportent des risques : morcellement éthique et *ethic shopping* en cas d'autonomisation ou compétition accrue entre concepts en cas d'intégration.

Cette étude ouvre des perspectives pour comprendre l'IA et l'innovation responsable. La littérature éthique en IA évolue, nécessitant une actualisation régulière et une vision moins « occidentale ». Enfin, l'analyse souligne la nécessité d'assurer un équilibre entre les concepts pour favoriser leur « méta-performativité ». Ce constat ouvre des perspectives d'approfondissement prometteuses pour comprendre comment différents concepts peuvent être intégrés de manière efficace et équilibrée.

RÉFÉRENCES

ADADI, A., BERRADA, M. (2018), Peeking inside the Black-Box : A Survey on Explainable Artificial Intelligence (XAI), *IEEE Access*, 6, 52138-52160.

ADAMS, R. (2021), Can Artificial Intelligence be Decolonized ?, *Interdisciplinary Science Reviews*, 46(1-2), 176-197.

AGGERI, F. (2017), Qu'est-ce que la performativité peut apporter aux recherches en management et sur les organisations : mise en perspective théorique et cadre d'analyse, *M@n@gement*, 20(1), 28-69.

ANTONIADI, A. M., DU, Y., GUENDOUZ, Y., WEI, L., MAZO, C., BECKER, B. A., MOONEY, C. (2021), Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems : A Systematic Review, *Applied Sciences*, 11(11), 5088.

ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., GARCIA, S., GIL-LOPEZ, S., MOLINA, D., BENJAMINS, R., CHATILA, R., HERRERA, F. (2020), Explainable Artificial Intelligence (XAI) : Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, *Information Fusion*, 58, 82-115.

BAGARIC, M., SVILAR, J., BULL, M., HUNTER, D., STOBBS, N. (2022), The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System : Transparent and Fair Artificial Intelligence ?, *American Criminal Law Review*, 59, 95-148.

BECKERS, R., KWADE, Z., ZANCA, F. (2021), The EU Medical Device Regulation : Implications for Artificial Intelligence-Based Medical Device Software in Medical Physics, *Physica Medica*, 83, 1-8.

BELLAMY, R. K., DEY, K., HIND, M., HOFFMAN, S. C., HOUDE, S., KANNAN, K., LOHIA, P., MARTINO, J., MEHTA, S., MOJSILOVIĆ A., NAGAR, S., NATESAN RAMAMURTHY, K., RICHARDS, J., SAHA, D., SATTIGERI, P., SINGH, M., VARSHNEY, K. R., ZHANG, Y. (2019), AI Fairness 360 : An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias, *IBM Journal of Research and Development*, 63(4/5), 4 :1-4 :15.

BENNETT, C. L., KEYES, O. (2020), What is the Point of Fairness ? Disability, AI and the Complexity of Justice, *ACM SIGACCESS Accessibility and Computing*, (125), 1-1.

BENZIDIA, S., MAKAOUI, N., BENTAHAR, O. (2021), The Impact of Big Data Analytics and Artificial Intelligence on Green Supply Chain Process Integration and Hospital Environmental Performance, *Technological Forecasting and Social Change*, 165, 120557.

BERENDT, B. (2019), AI for the Common Good ? ! Pitfalls, Challenges, and Ethics Pen-Testing ?, *Paladyn, Journal of Behavioral Robotics*, 10(1), 44-65.

BIETTI, E. (2021), From Ethics Washing to Ethics Bashing : A Moral Philosophy View on Tech Ethics, *Journal of Social Computing*, 2(3), 266-283.

BIRCHALL, C. (2011), Introduction to 'Secrecy and Transparency' : The Politics of Opacity and Openness, *Theory, Culture & Society*, 28(7-8), 7-25.

BIRD, S., DUDÍK, M., EDGAR, R., HORN, B., LUTZ, R., MILAN, V., SAMEKI, M., WALLACH, H., WALKER, K. (2020), Fairlearn : A Toolkit for Assessing and Improving Fairness in AI, *Microsoft, Tech. Rep.* MSR-TR-2020-32.

BONDI, E., XU, L., ACOSTA-NAVAS, D., KILLIAN, J. A. (2021), Envisioning Communities : A Participatory Approach towards AI for Social Good, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425-436.

BOURCIER, D., DE FILIPPI, P. (2018), Transparence des algorithmes face à l'Open Data : Quel statut pour les données d'apprentissage ?, *Revue française d'administration publique*, 167(3), 525-537.

BRAYNE, S., CHRISTIN, A. (2021), Technologies of Crime Prediction : The Reception of Algorithms in Policing and Criminal Courts, *Social Problems*, 68(3), 608-624.

BREVINI, B. (2020), Black Boxes, Not Green : Mythologizing Artificial Intelligence and Omitting the Environment, *Big Data & Society*, 7(2), 2053951720935141.

BRUNDAGE, M. (2016), Artificial Intelligence and Responsible Innovation, in *Fundamental Issues of Artificial Intelligence*, Cham, Springer, 543-554.

BRYNJOLFSSON, E., MITCHELL, T. (2017), What Can Machine Learning Do ? Workforce Implications, *Science*, 358(6370), 1530-1534.

BUHMANN, A., FIESELER, C. (2021), Towards a Deliberative Framework for Responsible Innovation in Artificial Intelligence, *Technology in Society*, 64, 101475.

BUHMANN, A., FIESELER, C. (2022), Deep Learning Meets Deep Democracy : Deliberative Governance and Responsible Innovation in Artificial Intelligence, *Business Ethics Quarterly*, 33(1), 146-179.

CABANTOUS, L., GOND, J. P. (2011), Rational Decision Making as Performative Praxis : Explaining Rationality's Éternel Retour, *Organization Science*, 22(3), 573-586.

CASTETS-RENARD, C. (2018), Régulation des algorithmes et gouvernance du machine learning : vers une transparence et « explicabilité » des décisions algorithmiques ? (Algorithm Regulation and Machine Learning Governance : Towards Transparency and 'Explainability' of Algorithmic Decisions ?), *Revue Droit & Affaires, Revue Paris II Assas*, 15e édition.

COHEN, I. G., EVGENIOU, T., GERKE, S., MINSSEN, T. (2020), The European Artificial Intelligence Strategy : Implications and Challenges for Digital Health, *The Lancet Digital Health*, 2(7), e376-e379.

COHEN, T., STILGOE, J., CAVOLI, C. (2018), Reframing the Governance of Automotive Automation : Insights from UK Stakeholder Workshops, *Journal of Responsible Innovation*, 5(3), 257-279.

COWLS, J. (2021), 'AI for Social Good' : Whose Good and Who's Good ? Introduction to the Special Issue on Artificial Intelligence for Social Good, *Philosophy & Technology*, 34(1), 1-5.

COWLS, J., TSAMADOS, A., TADDEO, M., FLORIDI, L. (2021), A Definition, Benchmark and Database of AI for Social Good Initiatives, *Nature Machine Intelligence*, 3(2), 111-115.

DABIC, M., MARZI, G., VLACIC, B., DAIM, T. U., VANHAVERBEKE, W. (2021), 40 Years of Excellence : An Overview of *Technovation* and a Roadmap for Future Research, *Technovation*, 106, 102303.

DANESHJOU, R., SMITH, M. P., SUN, M. D., ROTEMBERG, V., ZOU, J. (2021), Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms : A Scoping Review, *JAMA dermatology*, 157(11), 1362-1369.

DAUVERGE, P. (2020), Is Artificial Intelligence Greening Global Supply Chains ? Exposing the Political Economy of Environmental Costs, *Review of International Political Economy*, 29(3), 696-718.

DEBREF, R., GALLAUD, D., TEMPLE, L., TEMRI, L. (2019), Éditorial. L'innovation responsable, dimension stratégique des organisations, *Innovations*, 59(2), 5-13.

DIAKOPOULOS, N. (2020), Transparency, in *The Oxford Handbook of Ethics of AI*, 197-213.

DIGNUM, V. (2021), The Myth of Complete AI-fairness, in *Artificial Intelligence in Medicine : 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event*, June 15–18, Proceedings, Cham, Springer International Publishing, 3-8.

DILHAC, M. A., ABRASSART, C., VOARINO, N. (2018), *Rapport de la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*, Université de Montréal, 318 p.

EDWARDS, L., VEALE, M. (2017), Slave to the Algorithm : Why a Right to an Explanation is probably not the Remedy You Are Looking For, *Duke Law & Technology Review*, 16, 18.

FELZMANN, H., FOSCH-VILLARONGA, E., LUTZ, C., TAMÒ-LARRIEUX, A. (2020), Towards Transparency by Design for Artificial Intelligence, *Science and Engineering Ethics*, 26(6), 3333-3361.

FELZMANN, H., VILLARONGA, E. F., LUTZ, C., TAMO-LARRIEUX, A. (2019), Transparency You can Trust : Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns, *Big Data & Society*, 6(1), 2053951719860542.

FEUERRIEGEL, S., DOLATA, M., SCHWABE, G. (2020), Fair AI : Challenges and Opportunities, *Business & Information Systems Engineering*, 62(4), 379-384.

FLORIDI, L. (2021), The European Legislation on AI : A Brief Analysis of its Philosophical Approach, *Philosophy & Technology*, 34(2), 215-222.

FLORIDI, L., COWLS, J., BELTRAMETTI, M., CHATILA, R., CHAZERAND, P., DIGNUM, V., LUETGE, C., MADELIN, R., PAGALLO, U., ROSSI, F., SCHAFER, B., VALCKE, P., VAYENA, E. (2018), AI4People : An Ethical Framework for a Good AI Society : Opportunities, Risks, Principles, and Recommendations, *Minds and Machines*, 28(4), 689-707.

FLORIDI, L. (2019), Establishing the Rules for Building Trustworthy AI, *Nature Machine Intelligence*, 1(6), 261-262.

FLORIDI, L., COWLS, J., KING, T. C., TADDEO, M. (2020), How to Design AI for Social Good : Seven Essential Factors, *Science and Engineering Ethics*, 26(3), 1771-1796.

FOFFANO, F., SCANTAMBURLO, T., CORTÉS, A. (2022), Investing in AI for Social Good : An Analysis of European National Strategies, *AI & Society*, 38, 479-500.

FOURNIER-TOMBS, E. (2021), Towards a United Nations Internal Regulation for Artificial Intelligence, *Big Data & Society*, 8(2), 20539517211039493.

GALAZ, V., CENTENO, M. A., CALLAHAN, P. W., CAUSEVIC, A., PATTERSON, T., BRASS, I., BAUM, S., FARBER, D., FISCHER, J., GARCIA, D., MCPHEARSON, T., JIMENEZ, D., KING, B., LARCEY, P., LEVY, K. (2021), Artificial Intelligence, Systemic Risks, and Sustainability, *Technology in Society*, 67, 101741.

GAILHOFER, P., HEROLD, A., SCHEMMEL, J. P., SCHERF, C. S., DE STEBELSKI, C. U., KÖHLER, A. R., BRAUNGARDT, S. (2021), *The Role of Artificial Intelligence in the European Green Deal*, Luxembourg, Belgium, European Parliament.

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. (2016), *Deep Learning*, MIT Press.

GRINBAUM, A. (2018), Chance as a Value for Artificial Intelligence, *Journal of Responsible Innovation*, 5(3), 353-360.

GRØNSUND, T., AANESTAD, M. (2020), Augmenting the Algorithm : Emerging Human-in-the-Loop Work Configurations, *The Journal of Strategic Information Systems*, 29(2), 101614.

GUNNING, D. (2017), Explainable Artificial Intelligence (XAI), *Defence Advanced Research Projects Agency (DARPA)*, 2017/11.

GUNNING, D., STEFIK, M., CHOI, J., MILLER, T., STUMPF, S., YANG, G. Z. (2019), XA : Explainable Artificial Intelligence, *Science Robotics*, 4(37), eaay7120.

HAGENDORFF, T. (2020), The Ethics of AI Ethics : An Evaluation of Guidelines, *Minds and Machines*, 30(1), 99-120.

HANSEN, H. K., FLYVERBOM, M. (2015), The Politics of Transparency and The Calibration of Knowledge in the Digital Age, *Organization*, 22(6), 872-889.

HENDERSON, P., HU, J., ROMOFF, J., BRUNSKILL, E., JURAFSKY, D., PINEAU, J. (2020), Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning, *The Journal of Machine Learning Research*, 21(1), 10039-10081.

HERMANN, E. (2021), Leveraging Artificial Intelligence in Marketing for Social Good : An Ethical Perspective, *Journal of Business Ethics*, 179(1), 43-61.

HOLZINGER, A., DEHMER, M., EMMERT-STREIB, F., CUCCHIARA, R., AUGENSTEIN, I., DEL SER, J., SAMEK, W., JURISICA, I., DÍAZ-RODRÍGUEZ, N. (2022), Information Fusion as an Integrative Cross-Cutting Enabler to Achieve Robust, Explainable, and Trustworthy Medical Artificial Intelligence, *Information Fusion*, 79, 263-278.

HOLZMEYER, C. (2021), Beyond 'AI for Social Good' (AI4SG) : Social Transformations : not Tech-Fixes for Health Equity, *Interdisciplinary Science Reviews*, 46(1-2), 94-125.

JOBIN, A., IENCA, M., VAYENA, E. (2019), The Global Landscape of AI Ethics Guidelines, *Nature Machine Intelligence*, 1, 389-399.

JOHN-MATHEWS, J. M., CARDON, D., BALAGUÉ, C. (2022), From Reality to World : A Critical Perspective on AI Fairness, *Journal of Business Ethics*, 178(4), 945-959.

KAPLAN, A. M., HAENLEIN, M. (2019), Siri, Siri, in my Hand : Who's The Fairest in The Land ? On the Interpretations, Illustrations, and Implications of Artificial Intelligence, *Business Horizons*, 62(1), 15-25.

KHAREGHANI, S. (2020), Capitalizing on AI's Potential to Help Tackle the Climate Crisis [Opinion], *IEEE Technology and Society Magazine*, 39(2), 41-47.

KELLOGG, K. C., VALENTINE, M. A., CHRISTIN, A. (2020), Algorithms at Work : The New Contested Terrain of Control, *Academy of Management Annals*, 14(1), 366-410.

KORMELINK, G. (2019), *Responsible Innovation, Ethics, Safety and Technology : How to Deal with Risks and Ethical Questions Raised by the Development of New Technologies* (2nd ed.), Delft, TU Delft Open

LANNELONGUE, L., GREALEY, J., INOUYE, M. (2021), Green Algorithms : Quantifying the Carbon Footprint of Computation, *Advanced Science*, 8(12), 2100707.

LANDERS, R. N., BEHREND, T. S. (2022), Auditing the AI Auditors : A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models, *American Psychologist*, 78(1), 36-49.

LARSSON, S., HEINTZ, F. (2020), Transparency in Artificial Intelligence, *Internet Policy Review*, 9(2).

LILKOV, D. (2021), Regulating Artificial Intelligence in the EU : A Risky Game, *European View*, 20(2), 166-174.

LINKOV, I., TRUMP, B. D., POINSATTE-JONES, K., FLORIN, M. V. (2018), Governance Strategies for a Sustainable Digital World, *Sustainability*, 10(2), 440.

LIPTON, Z. C. (2018), The Mythos of Model Interpretability, *Commun ACM 2018 ; 61*, 36–43.

LIU, R., JIA, C., WEI, J., XU, G., VOSOUGHI, S. (2022), Quantifying and Alleviating Political Bias in Language Models, *Artificial Intelligence*, 304, 103654.

MACKENZIE, D., MILLO, Y. (2003), Construction d'un marché et performation théorique : Sociologie historique d'une bourse de produits dérivés financiers, *Réseaux*, 6, 15-61.

MACKENZIE, D., MUNIESA, F., SIU, L. (2007), *Do Economists Make Markets ? On the Performativity of Economics*, Princeton University Press.

MCCLURE, P. K. (2018), "You're Fired," Says the Robot : The Rise of Automation in The Workplace, Technophobes, and Fears of Unemployment, *Social Science Computer Review*, 36(2), 139-156.

MADAIO, M., EGEDE, L., SUBRAMONYAM, H., WORTMAN VAUGHAN, J., WALLACH, H. (2022), Assessing the Fairness of AI Systems : AI Practitioners' Processes, Challenges, and Needs for Support, *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 1-26.

MAKRIDAKIS, S. (2017), The Forthcoming Artificial Intelligence (AI) Revolution : Its Impact on Society and Firms, *Futures*, 90, 46-60.

MARTI, E., GOND, J. P. (2018), When Do Theories Become Self-Fulfilling ? Exploring the Boundary Conditions of Performativity, *Academy of Management Review*, 43(3), 487-508.

MARTIN, K. (2019), Ethical Implications and Accountability of Algorithms, *Journal of Business Ethics*, 160(4), 835-850.

MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., GALSTYAN, A. (2021), A Survey on Bias and Fairness in Machine Learning, *ACM Computing Surveys (CSUR)*, 54(6), 1-35.

MÉNISSIER, T. (2020), A "Machiavellian Moment" for Artificial Intelligence ? The Montreal Declaration for the Responsible Development of AI, *Raisons Politiques*, 77(1), 67-81.

MENSAH, J. (2019), Sustainable Development : Meaning, History, Principles, Pillars, and Implications for Human Action : Literature Review, *Cogent Social Sciences*, 5(1), 1653531.

MERHI, M. I. (2022), An Assessment of the Barriers Impacting Responsible Artificial Intelligence, *Information Systems Frontiers*, 1-14.

MESKE, C., BUNDE, E., SCHNEIDER, J., GERSCH, M. (2022), Explainable Artificial Intelligence : Objectives, Stakeholders, and Future Research Opportunities, *Information Systems Management*, 39(1), 53-63.

MITTELSTADT, B. (2019), Principles Alone cannot Guarantee Ethical AI, *Nature Machine Intelligence*, 1(11), 501-507.

MITTELSTADT, B. D., ALLO, P., TADDEO, M., WACHTER, S., FLORIDI, L. (2016), The Ethics of Algorithms : Mapping the Debate, *Big Data & Society*, 3(2), 2053951716679679.

MÖKANDER, J., AXENTE, M., CASOLARI, F., FLORIDI, L. (2022), Conformity Assessments and Post-Market Monitoring : A Guide to the Role of Auditing in the Proposed European AI Regulation, *Minds and Machines*, 32(2), 241-268.

MOLINA RODRÍGUEZ-NAVAS, P., MEDRANDA MORALES, N., MUÑOZ LALINDE, J. (2021), Transparency for Participation through the Communication Approach, *ISPRS International Journal of Geo-Information*, 10(9), 586.

MOLNAR, C., CASALICCHIO, G., BISCHL, B. (2020), Interpretable Machine Learning : A Brief History, State-of-the-Art and Challenges, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Cham, Springer, 417-431.

MORA-CANTALLOPS, M., SANCHEZ-ALONSO, S., GARCIA-BARRIOCANAL, E., SICILIA, M. A. (2021), Traceability for Trustworthy AI : A Review of Models and Tools, *Big Data and Cognitive Computing*, 5(2), 20.

MUNOKO, I., BROWN-LIBURD, H. L., VASARHELYI, M. (2020), The Ethical Implications of Using Artificial Intelligence in Auditing, *Journal of Business Ethics*, 167(2), 209-234.

MURDOCK, G., BREVINI, B. (2019), Communications and the Capitalocene : Disputed Ecologies, Contested Economies, Competing Futures, *The Political Economy of Communication*, 7(1).

NTOUTSI, E., FAFALIOS, P., GADIRAJU, U., IOSIFIDIS, V., NEJDL, W., VIDAL, M. E., RUGGIERI, S., TURINI, F., PAPADOPOULOS, S., KRASANAKIS, E., KOMPATSIARIS, I., KINDER-KURLANDA, K., WAGNER, C., KARIMI, F., FERNANDEZ, M., ALANI, H., BERENDT, B., KRUEGEL, T., HEINZE, C., BROELEMANN, K., KASNECI, G., TIROPANIS, T, STAAB, S. (2020), Bias

in DataŒDriven Artificial Intelligence Systems : An Introductory Survey, *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 10(3), e1356.

OSTROM, E. (1990), *Governing the Commons : The Evolution of Institutions for Collective Action*, Cambridge University Press.

OWENS, K., WALKER, A. (2020), Those Designing Healthcare Algorithms Must Become Actively Anti-Racist, *Nature Medicine*, 26(9), 1327-1328.

PARLY, F. (2019), Avant-propos – Intelligence artificielle et défense, *Revue défense nationale*, 820(5), 9-17.

PASQUALE, F. (2015), *The Black Box Society : The Secret Algorithms that Control Money and Information*, Harvard University Press.

PRUNKL, C. E., ASHURST, C., ANDERLJUNG, M., WEBB, H., LEIKE, J., DAFOE, A. (2021), Institutionalizing Ethics in AI through Broader Impact Requirements, *Nature Machine Intelligence*, 3(2), 104-110.

PWC (2019), *How AI can Enable a Sustainable Future*, PWC Insights.

RAMBACH, P. (2022), Business Transformation towards Sustainability : Embracing AI at Scale, *Schneider Electrics Blog*, 22(06). Retrieved from : https://blog.se.com/digital-transformation/internet-of-things/2022/05/13/embracing-ai-at-scale/

RESSÉGUIER, A., RODRIGUES, R. (2020), AI Ethics should not Remain Toothless ! A Call to Bring Back the Teeth of Ethics, *Big Data, & Society*, 7(2), 2053951720942541.

RIBEIRO, B. E., SMITH, R. D., MILLAR, K. (2017), A Mobilising Concept ? Unpacking Academic Representations of Responsible Research and Innovation, *Science and Engineering Ethics*, 23, 81-103.

RIEDER, G., SIMON, J., WONG, P. H. (2021), Mapping the Stony Road toward Trustworthy AI, *Machines We Trust : Perspectives on Dependable AI*, 27.

ROBERT, L. P., PIERCE, C., MARQUIS, L., KIM, S., ALAHMAD, R. (2020), Designing Fair AI for Managing Employees in Organizations : A Review, Critique, and Design Agenda, *Human–Computer Interaction*, 35(5-6), 545-575.

ROBINSON, S. C. (2020), Trust, Transparency, and Openness : How Inclusion of Cultural Values Shapes Nordic National Public Policy Strategies for Artificial Intelligence (AI), *Technology in Society*, 63, 101421.

ROLNICK, D., DONTI, P. L., KAACK, L. H., KOCHANSKI, K., LACOSTE, A., SANKARAN, K., ROSS, A. S., MILOJEVIC-DUPONT, N., JAQUES, N., WALDMAN-BROWN, A., LUCCIONI, A., MAHARAJ, T., SHERWIN, E. D., MUKKAVILLI, S. K., KORDING, K. P., GOMES, C., NG, A. Y., HASSABIS, D., PLATT, J. C., CREUTZIG, F., CHAYES, J., BENGIO, Y. (2019), Tackling Climate Change with Machine Learning, *ACM Computing Surveys (CSUR)*, 55(2), 1-96.

SANTONI DE SIO, F., MECACCI, G. (2021), Four Responsibility Gaps with Artificial Intelligence : Why They Matter and How to Address Them, *Philosophy & Technology*, 34, 1057-1084.

SCHWARTZ, R., DODGE, J., SMITH, N. A., ETZIONI, O. (2020), Green AI, *Communications of the ACM*, 63(12), 54-63.

SHABAN-NEJAD, A., MICHALOWSKI, M., BUCKERIDGE, D. L. (2021), Explainability and Interpretability : Keys to Deep Medicine, in *Explainable AI in Healthcare and Medicine : Building a Culture of Transparency and Accountability*, Cham, Springer, 1-10.

SHRESTHA, Y. R., BEN-MENAHEM, S. M., VON KROGH, G. (2019), Organizational Decision-Making Structures in the Age of Artificial Intelligence, *California Management Review*, 61(4), 66-83.

SINGH, S., SHARMA, P. K., YOON, B., SHOJAFAR, M., CHO, G. H., RA, I. H. (2020), Convergence of Blockchain and Artificial Intelligence in IoT Network for the Sustainable Smart City, *Sustainable Cities and Society*, 63, 102364.

STAHL, B. C. (2013), Responsible research and innovation : The role of privacy in an emerging framework. *Science and Public* Policy, 40(6), 708-716.

STAHL, B. C. (2021), Concepts of Ethics and Their Application to AI, in *Artificial Intelligence for a Better Future*, Cham, Springer, 19-33.

STAHL, B. C. (2022), Responsible Innovation Ecosystems : Ethical Implications of the Application of the Ecosystem Concept to Artificial Intelligence, *International Journal of Information Management*, 62, 102441.

STAHL, B. C., WRIGHT, D. (2018), Ethics and Privacy in AI and Big Data : Implementing Responsible Research and Innovation, *IEEE Security and Privacy*, 16(3), 26-33.

STILGOE, J., OWEN, R., MACNAGHTEN, P. (2013), Developing a Framework for Responsible Innovation, *Research Policy*, 42(9), 1568-1580.

TADDEO, M., FLORIDI, L. (2018), How AI Can Be a Force for Good, *Science*, 361(6404), 751-752.

TANG, B. W. (2020), Independent AI Ethics Committees and ESG Corporate Reporting on AI as Emerging Corporate and AI Governance Trends, in Chishti, S., Batorletti, I., Leslie, A., Millie, S. M. (eds), *The AI Book : The Artificial Intelligence Handbook for Investors, Entrepreneurs and FinTech Visionaries*, 180-185.

TE KULVE, H., RIP, A. (2011), Constructing Productive Engagement : Pre-Engagement Tools for Emerging Technologies, *Science and Engineering Ethics*, 17, 699-714.

TIELL, S. (2019), Create an Ethics Committee to Keep Your AI Initiative in Check, *Harvard Business Review*, 15.

TOORAJIPOUR, R., SOHRABPOUR, V., NAZARPOUR, A., OGHAZI, P., FISCHL, M., (2021), Artificial Intelligence in Supply Chain Management : A Systematic Literature Review, *Journal of Business Research*, 122, 502-517.

UMBRELLO, S., VAN DE POEL, I. (2021), Mapping Value Sensitive Design onto AI for Social Good Principles, *AI and Ethics*, 1(3), 283-296.

VAKKURI, V., KEMELL, K. K., KULTANEN, J., SIPONEN, M., ABRAHAMSSON, P. (2022), Ethically Aligned Design of Autonomous Systems : Industry Viewpoint and an Empirical Study, *EJBO – Electronic Journal of Business Ethics and Organization Studies*, 1(27), 4-15.

VAN ECK, N., WALTMAN, L. (2010), Software Survey : Vosviewer, a Computer Program for Bibliometric Mapping, *Scientometrics*, 84(2), 523-538.

VAN NOOD, R., YEOMANS, C. (2021), Fairness as Equal Concession : Critical Remarks on Fair AI, *Science and Engineering Ethics*, 27(6), 1-14.

VAN WYNSBERGHE, A. (2021), Sustainable AI : AI for Sustainability and the Sustainability of AI, *AI and Ethics*, 1(3), 213-218.

VEALE, M., BORGESIUS, F. Z. (2021), Demystifying the Draft EU Artificial Intelligence Act : Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach, *Computer Law Review International*, 22(4), 97-112.

VINUESA, R., AZIZPOUR, H., LEITE, I., BALAAM, M., DIGNUM, V., DOMISCH, S., FELLÄNDER, A., LANGHANS, S. D., TEGMARK, M., FUSO NERINI, F. (2020), The Role of Artificial Intelligence in Achieving the Sustainable Development Goals, *Nature Communications*, 11(1), 233.

VON ESCHENBACH, W. J. (2021), Transparency and The Black Box Problem : Why We Do Not Trust AI, *Philosophy & Technology*, 34(4), 1607-1622.

WACHTER, S., MITTELSTADT, B., FLORIDI, L. (2017), Transparent, Explainable, and Accountable AI for Robotics, *Science Robotics*, 2(6), eaan6080.

WACHTER, S., MITTELSTADT, B., RUSSELL, C. (2021), Why Fairness Cannot Be Automated : Bridging the Gap Between EU Non-Discrimination Law and AI, *Computer Law & Security Review*, 41, 105567.

WAGNER, B. (2018), Ethics as an Escape from Regulation : From "Ethics-Washing" to Ethics-Shopping ?, in Bayamliogu, E., Baraliuc, I., Janssens, L.A.W., Hildebrandt, M. (eds), *Being Profiled : Cogitas Ergo Sum. 10 Years of 'Profiling the European Citizen'*, Amsterdam, Amsterdam University Press, 84-88.

WALMSLEY, J. (2021), Artificial Intelligence and the Value of Transparency, *AI & Society*, 36(2), 585-595.

WAMBA, S. F., BAWACK, R. E., GUTHRIE, C., QUEIROZ, M. M., CARILLO, K. D. A. (2021), Are We Preparing for a Good AI Society ? A Bibliometric Review and Research Agenda, *Technological Forecasting and Social Change*, 164, 120482.

WASHINGTON, A. L. (2018), How to Argue with An Algorithm : Lessons from the COMPAS-Propublica Debate, *The Colorado Technology Law Journal*, 17, 131.

WATSON, D. S., FLORIDI, L. (2021), The Explanation Game : A Formal Framework for Interpretable Machine Learning, in *Ethics, Governance, and Policies in Artificial Intelligence*, Cham, Springer International Publishing, 185-219.

WEAVER, J. F. (2018), Everything is not Terminator : America's First AI Legislation, *Journal of Robotics Artificial Intelligence and Law (RAIL)*, 1(3), 201-207.

WEHBE, R. M., SHENG, J., DUTTA, S., CHAI, S., DRAVID, A., BARUTCU, S., WU, Y., CANTRELL, D. R., XIAO, N., ALLEN, B. D., MACNEALY, G. A., SAVAS, H., ARGAWAL, R., PAREKH, N., KATSAGGELOS, A. K. (2021), DeepCOVID-XR : An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large US Clinical Data Set, *Radiology*, 299(1), E167-E176.

WONG, D. (2018), VOSviewer, *Technical Services Quarterly*, 35(2), 219-220.

WOSK, J. (2010), Metropolis, *Technology and Culture*, 51(2), 403-408.

WU, C.-J., RAGHAVENDRA, R., GUPTA, U., ACUN, B., ARDALANI, N., MAENG, K., CHANG, G., AGA BEHRAM, F., HUANG, J., BAI, C., GSCHWIND, M., GUPTA, A., OTT, M., MELNIKOV, A., CANDIDO, S., BROOKS, D. C., CHAUHAN, G., LEE, B., LEE, H.-H., AKYILDIZ, B., BALANDAT, M., SPISAK, J., JAIN, R., RABBAT, M., HAZELWOOD, K. (2022), Sustainable AI : Environmental Implications, Challenges and Opportunities, *Proceedings of Machine Learning and Systems*, 4, 795-813.

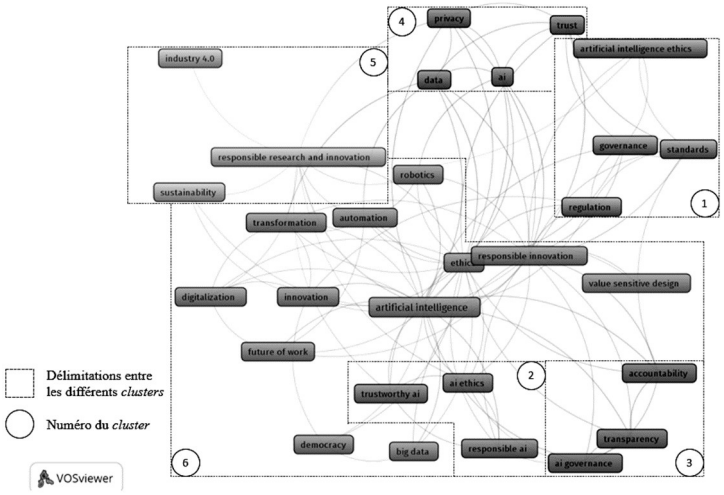
ZANZOTTO, F. M. (2019), Human-in-the-Loop Artificial Intelligence, *Journal of Artificial Intelligence Research*, 64, 243-252.

ZEDNIK, C. (2021), Solving the Black Box Problem : A Normative Framework for Explainable Artificial Intelligence, *Philosophy & Technology*, 34(2), 265-288.

ZIELKE, T. (2020), Is Artificial Intelligence Ready for Standardization ?, in *European Conference on Software Process Improvement*, Cham, Springer, 259-274.

Annexes

Figure 3 – Clusterisation par co-occurrences thématiques sur Vosviewer



Le graphique suivant est réalisé via le logiciel Vosviewer (*cf.* Van Eck, Waltman, 2010 ; Wong, 2018 ; Dabić *et al.*, 2021) à partir d'un corpus de 207 articles collectés sur la plateforme Scopus répondant au critère de mots clés suivant : « artificial intelligence » and « responsible innovation ».

Tableau 2 – Thèmes des travaux inclus dans la revue de littérature

N° du cluster	Thèmes des travaux inclus dans la revue de littérature
1	Travaux sur l'articulation entre éthique de l'IA et gouvernance
2	Travaux portant sur la capacité à élaborer des règles permettant d'articuler toutes les parties prenantes, afin de construire une confiance réciproque entre acteurs de l'IA et avec les utilisateurs
3	Travaux qui se focalisent principalement sur la question de l'*accountability* (rendu de compte) et la transparence
4	Travaux sur le management et la gouvernance des données et les exigences en termes de consentement et de respect de la vie privée
5	Travaux sur la transformation des processus de production, et leur soutenabilité
6	Travaux sur les conséquences induites par l'IA sur la société et les systèmes socio-productifs et démocratiques, et les risques, mais aussi les opportunités offertes par cette technologie à condition de s'inscrire dans des démarches de type innovation responsable

En lien avec la figure 3, le tableau 2 présente les différents thèmes qui ressortent de la *clusterisation* par co-occurrences thématiques sur Vosviewer. Le premier sous-groupe correspond aux *clusters* 1, 2 et 3 et traite des prérequis éthiques lors du processus de conception, incluant les enjeux de régulation et de gouvernance. Le deuxième sous-groupe correspond aux concepts du *cluster* numéro 5 et se concentre sur les changements induits par l'IA dans les systèmes productifs, en particulier sur le plan industriel, et la question de la soutenabilité de ces transformations. Enfin, le troisième sous-groupe inclut les concepts du *cluster* 6 et examine les conséquences de l'IA et la capacité de l'innovation responsable à influencer son développement dans le but de servir le bien commun. Les travaux ressortant du *cluster* 4 sur le management et la gouvernance des données ne concernent pas directement des concepts d'IA éthique mais alimentent les réflexions autour de ces derniers.