Transparence et explicabilité de l'IA : enjeux et demandes sociales dans le cas des industries culturelles

Article inédit, mis en ligne le 20 avril 2026

Jaércio da Silva

Maître de conférences en sciences de l'information et de la communication à l'université Paris-Panthéon-Assas (IFP/CARISM)
jaercio.dasilva@assas-universite.fr

Joëlle Farchy

Professeure en sciences de l'information et de la communication à l'université Paris 1 Panthéon-Sorbonne. Directrice de la Chaire Pluralisme culture et éthiques du numérique (EMNS/PcEn)
joelle.farchy@univ-paris1.fr

Plan de l'article

RÉSUMÉ

L'essor de l'intelligence artificielle suscite des débats croissants sur la transparence et l'explicabilité, souvent invoquées sous l'angle de l'éthique de l'IA, mais dont la mise en œuvre demeure un défi technique et politique. Cet article analyse ces exigences dans les industries culturelles, où l'IA intervient dans la production, la recommandation et la création de contenus. Après une exploration du champ de l'XAI, visant à rendre l'IA plus

compréhensible, nous examinons les tensions entre automatisation et intervention humaine. Enfin, nous mettons en lumière les régulations émergentes, notamment le règlement européen sur l'IA, qui impose des obligations de transparence et soulève des questions sur le droit d'auteur et le partage de la valeur.

Mots clés

Intelligence artificielle, explicabilité, industries culturelles, XAI, droit d'auteur

Title

AI transparency and explainability: challenges and social demands in the case of cultural industries

Abstract

The rise of artificial intelligence has sparked increasing debates on transparency and explainability, often framed within AI ethics but challenging to implement both technically and politically. This article examines how these requirements manifest in cultural industries, where AI plays a key role in content production, recommendation, and creation. First, we explore the field of XAI, which aims to make AI systems more interpretable. We then analyze the tensions between automation and human intervention in cultural industries. Finally, we highlight emerging regulatory frameworks, particularly the European AI Act, which imposes transparency obligations on companies and raises critical questions about copyright protection and value distribution.

Keywords

Artificial intelligence, explainability, cultural industries, XAI, copyright

TÍTULO

Transparencia y explicabilidad de la IA : desafíos y demandas sociales en el caso de las industrias culturales

Resumen

El auge de la inteligencia artificial ha generado un creciente debate sobre la transparencia y la explicabilidad, conceptos frecuentemente enmarcados dentro de la ética de la IA, pero cuya implementación sigue representando un desafío tanto técnico como político. Este artículo examina cómo estos requisitos se manifiestan en las industrias culturales, donde la IA desempeña un papel fundamental en la producción, recomendación y creación de contenidos. En primer lugar, exploramos el campo de la XAI, cuyo objetivo es hacer que los sistemas de IA sean más comprensibles. Posteriormente, analizamos las tensiones entre la automatización y la intervención humana en estas industrias. Finalmente, destacamos los marcos regulatorios emergentes, en particular el Reglamento Europeo sobre IA, que impone obligaciones de transparencia a las empresas y plantea cuestiones clave sobre la protección de los derechos de autor y la distribución del valor.

Palabras clave

Inteligencia artificial, explicabilidad, industrias culturales, XAI, derechos de autor

INTRODUCTION

L'intelligence artificielle (IA) ne relève plus uniquement du domaine scientifique ou technique : elle s'est imposée comme un objet central du débat public, suscitant l'attention croissante des décideurs politiques, des professionnels, des journalistes, des citoyens et des experts. Sa mise en visibilité médiatique participe de sa construction en tant que problème public (Chateauraynaud, 2019), et souligne son rôle stratégique dans les dynamiques de pouvoir économique et géopolitique contemporaines (Kirtchik, 2019 ; Crawford, 2021). À travers l'analyse des discours dans la presse d'information générale anglophone, Crépel et Cardon (2022) démontrent la coexistence de deux grands registres d'appréhension de la question : l'un, fondé sur l'idée de remplacement, projette l'IA comme une menace existentielle pour l'humanité ; l'autre, d'inspiration légaliste, plaide pour un encadrement de la relation homme-machine. Mais au-delà des préoccupations qui réduisent l'IA à ses effets de pouvoir et d'impact, divers questionnements émergent au fil du développement de cette innovation, qui ont trait aux notions de transparence et d'explicabilité de cette technologie. Pour mieux cerner la manière dont se posent ces enjeux, et par qui ils sont posés, nous faisons une lecture analytique d'une littérature en plein essor : celle de l'*Explainable Artificial Intelligence* (Van Lent *et al.*, 2004). Ce champ nous permet d'interroger les tensions entre innovation technologique, stratégies industrielles et politiques de régulation, en les observant à travers le prisme des industries culturelles, secteur déjà fortement traversé par l'automatisation de tâches et leur délégation à des entités artificielles (Attencourt *et al.*, 2025), qu'elles soient décrites comme algorithmiques, génératives ou robotiques.

La mise en débat public de l'IA et sa médiatisation en France sont souvent regroupées sous l'étiquette « éthique de l'IA » (Beaudouin et Velkovska, 2023). La réflexion éthique sur les développements des IA s'enracine d'abord dans le monde scientifique à partir des années 2000. L'« éthique de l'IA » n'est alors pas pensée comme un outil de régulation à destination des pouvoirs publics, mais, au contraire, son développement vise à contrer une potentielle mainmise du législateur sur l'activité scientifique. Le lancement de la tribune de l'organisation *The Future of Life Institute*[1] en 2015 marque un tournant médiatique : les journalistes inscrivent désormais des questions nouvelles à l'agenda des politiques publiques et contribuent à transformer l'« éthique de l'IA » d'un mode de réflexivité scientifique à une catégorie d'action publique et de régulation. La formule valise « éthique de l'IA » désigne donc désormais le pendant social de la technique, les aspects non techniques des systèmes algorithmiques, c'est-à-dire les interrogations sur leurs usages, leurs conséquences sociales ou et leurs régulations juridiques.

Le nombre de publications cherchant à établir des principes éthiques pour le développement de l'IA à la suite de directives émises par divers organismes (industries, gouvernements, organisations de la société civile, etc.) illustre le déploiement intensif des IA dans différents secteurs et domaines. Jobin, Ienca et Vayena (2019) ont analysé 84 rapports institutionnels contenant des principes éthiques ou des lignes directrices pour l'IA. Les chercheurs arrivent au constat que, bien qu'aucun principe éthique unique ne soit commun à tous les documents, une convergence apparaît autour de quelques notions, celles de « justice », « d'équité », de « non-malfaisance », de « responsabilité », de « confidentialité » ou encore de « transparence », soit la demande d'ouvrir la fameuse « boîte noire » des algorithmes.

La critique de l'IA et les discours éthiques sont également produits par les acteurs qui façonnent l'innovation (Beaudouin et Velkovska, 2023). Bélisle-Pipon *et al.* (2023) ont mis en évidence une contradiction : alors que les rapports institutionnels prônent la transparence

dans la conception des systèmes techniques, peu d'entre eux précisent comment ils ont été rédigés, qui y a participé (y compris les financeurs) ou quelle approche méthodologique a été adoptée. Le secteur privé est largement surreprésenté parmi les sources consultées lors de la rédaction de ces rapports. Un autre exemple est la lettre signée en 2015 par Elon Musk, Stephen Hawking, Steve Wozniak et Noam Chomsky, alertant sur les risques existentiels liés aux avancées de l'IA. Les entreprises sont donc prises entre plusieurs impératifs : contrôler la perception de leurs techniques, construire un récit autour de l'émergence imminente d'une IA super-humaine et, en parallèle, afficher une certaine transparence sur leurs pratiques.

La compréhension des objets techniques n'est certes pas une condition de la réussite d'une innovation et de son succès sur le marché comme l'a montré la sociologie des agencements marchands (Callon *et al.*, 2013). Nous ne comprenons souvent que partiellement le vaste réseau sociotechnique qui existe derrière le fonctionnement d'un dispositif (Callon, 2006). Le succès d'une innovation repose plutôt sur le consensus autour de ses qualifications et de sa singularité (Callon *et al*. 2000), ce qui permet d'établir une relation de confiance entre l'utilisateur et la technique. La question est désormais de savoir si l'on peut déléguer tout ou partie de nos activités à l'IA, et à quelles conditions. Ainsi, ce n'est pas tant le fonctionnement interne de la technologie qui importe, mais la réplicabilité d'un résultat obtenu selon un procédé identique. L'IA ne fait pas exception à l'histoire des techniques. L'opacité de fonctionnement n'a, a priori, pas freiné le succès de sa diffusion auprès du grand public comme le démontre le succès de ChatGPT : un million d'utilisateurs en cinq jours et 180 millions d'utilisateurs actifs mensuels[2]. On connait le « *privacy paradox* » (Barnes, 2006) qui illustre la capacité des usagers de services numériques à abandonner une partie de leur vie privée en fournissant des données personnelles aux plateformes, tout en étant inquiets de cet abandon. De manière similaire, nous pourrions dire qu'avec l'IA existe une sorte « d'*opacity paradox* », car les usagers adoptent ces outils, malgré des résultats que l'on pressent risqués, non maîtrisés, voire néfastes à l'environnent et au droit du travail. Reste alors à comprendre dans le cas de l'IA, qui sont les acteurs à l'origine d'une demande de transparence, quelles sont leurs motivations et dans quelle mesure ils apportent chacun leurs propres intérêts, questionnements et manières de faire émerger une parole publique sur le sujet. *In fine*, les demandes de transparence contribuent-elles à façonner le débat public sur l'IA ?

Cet article[3] ne repose pas sur un corpus empirique ou un travail de terrain. Il s'appuie sur un état de la littérature en XAI (Ribeiro *et al.*, 2016 ; Miller, 2019) pour interroger une injonction dans les discours des acteurs publics et privés : celle de la transparence. Présentée comme un principe « éthique » incontestable, cette notion est ici analysée de manière critique à travers une analyse structurée en quatre parties. Dans une première partie, nous étudions les diverses approches théoriques et méthodologiques proposées dans la littérature académique pour aborder la question de l'explicabilité, qu'il s'agisse d'étudier les mécanismes grâce auxquels le modèle fonctionne ou les contextes sociaux, culturels et matériels dans lesquels les pratiques prennent place. En examinant, en deuxième partie, le cas concret des industries culturelles dans lesquelles l'IA est déjà fortement implantée, nous montrons de quelle manière, au-delà du principe éthique général de transparence, se mettent en œuvre, de manière opérationnelle, des demandes techniques et sociales. Ces dernières concernent l'économie de la création, tant du point de vue des conditions de travail, qui seront l'objet de la partie 3, que de la régulation et du partage de la valeur, sujet particulièrement incarné dans le débat autour des droits d'auteur, abordé en partie 4.

LA RECHERCHE SCIENTIFIQUE AU SERVICE DE LA TRANSPARENCE

Il existe plusieurs manières d'expliquer le fonctionnement des IA, et des chercheurs se sont saisis de cette interrogation afin de proposer plusieurs approches. Une « science-carrefour » constituée en branches disparates, caractérisées par des approches, des temporalités et cheminements intellectuels différents émerge (Kirtchik, 2019). Ce champ de recherche se situe à la frontière entre l'académique et les industries du numérique, avec une approche multidisciplinaire qui rencontre dans le dialogue surtout communautaire sa principale limite.

Ce qu'on appelle l'XAI est un champ de recherche foisonnant qui regroupe une série de processus et de techniques visant à rendre les IA compréhensibles et explicables pour les humains et à leur donner la possibilité d'étudier ces IA pour comprendre leur mode de fonctionnement. L'objectif peut également être le développement d'agents virtuels comme outils d'explicabilité, dans le cas où le modèle s'auto-explique (Weitz *et al.*, 2021). Ce champ se complexifie au fil des années, en particulier avec le développement de sous-catégories de l'IA comme le *Machine Learning* (ML) et son sous-ensemble, le *Deep Learning*. Nous sommes ainsi passés de modèles d'IA faible (*Narrow AI*), conçus pour effectuer une tâche spécifique ou un ensemble limité de tâches (par exemple, un algorithme de recommandation sur une plateforme), à des modèles artificiels où les IA apprennent à partir de données pour établir des prédictions ou prendre des décisions sans être explicitement programmés pour chaque tâche, parfois en utilisant des réseaux de neurones artificiels pour modéliser des données (comme c'est le cas des IA génératives). Sans surprise, plus les systèmes sont complexes et autonomes, plus il est difficile d'expliquer leur fonctionnement.

En complément des travaux en informatique et en Interaction Humain-Machine (IHM) abordant cette problématique qui constituent le cœur de ce que l'on nomme l'XAI, des travaux en sciences humaines et sociales replacent la question dans un contexte global.

Favoriser la compréhension des humains

Les approches ne sont pas monolithiques et Adadi et Berrada (2018) démontrent l'existence de deux approches en XAI. Tout d'abord, une forme d'explicabilité « intra-modèle », où l'explicabilité est *by design*, avec des IA conçues intrinsèquement comme explicables, dans une approche globale où l'informaticien/utilisateur, peuvent expliquer le mode de fonctionnement d'un modèle entier. Deuxièmement, l'explicabilité comme interprétation des résultats obtenus post-modélisation, dans ce cas l'informaticien/utilisateur attribue une explication à partir de modèles préalablement construits et entraînés aux résultats qu'il obtient de la machine. Dans les deux cas, leur objectif principal est d'optimiser l'articulation entre la performance du modèle et son explicabilité par la prédiction mathématique, visant ainsi à améliorer et à mesurer l'efficacité de l'IA. Ils se concentrent sur la relation entre les entrées (*inputs*) et les sorties (*outputs*) dès la conception, ce qui a pour intérêt de prédire la stabilité du modèle en examinant la probabilité d'obtenir des résultats comparables lorsqu'on utilise le même modèle sur des jeux de données différents.

Afchar définit l'explicabilité pour les chercheurs en ML comme « un vecteur qui évalue les influences de certains éléments explicatifs dans les processus de calcul d'un résultat cible » (Afchar, 2023, p. 165). Ici, l'explication est une affaire qui peut être résolue par la probabilité mathématique. Afchar s'éloigne d'une approche méthodologique générale (Adadi et Berrada, 2018 ; Gunning, 2017 ; Rudin, 2019) pour proposer une étude de cas sur les systèmes de recommandation musicale (Afchar *et al.*, 2022). Il affirme que les explications en ML peuvent principalement être définies comme la recherche des scores d'influence des caractéristiques d'entrée (p. 196). Cela signifie que chaque caractéristique d'entrée d'un modèle a un poids ou une importance dans la prise de décision de la machine

et, partant, du résultat qu'elle produit. Comprendre ces scores d'influence donne la possibilité de vérifier si le modèle prend des décisions cohérentes et d'identifier des biais dans les décisions qui sont effectuées par le dispositif. Toutefois, Afchar souligne que la manière dont sont attribués des scores d'influence aux caractéristiques d'entrée peut varier d'un modèle à l'autre, rendant cette méthode arbitraire (Afchar, 2023).

Dans le cadre de l'IHM, l'explicabilité prend une dimension supplémentaire, que l'on désigne par l'explicabilité post-modélisation (*post hoc*) c'est-à-dire la capacité des développeurs (et des utilisateurs) à interpréter le résultat des dispositifs d'IA (Guidotti *et al.*, 2018 ; Doshi-Velez et Kim, 2017). Dans ce cas, l'utilisateur se voit expliquer un résultat particulier ou une décision spécifique de manière locale, c'est-à-dire propre à son contexte d'usage. L'explicabilité se concentre alors moins sur la compréhension interne du modèle que sur son résultat final (Lipton, 2018). Lipton préfère le terme *interpretability*, car selon lui, les prédictions et les métriques calculées intramodèles ne suffisent pas toujours à expliquer le modèle (*ibid.*, p. 57). Le terme « interpréter » apparait ainsi plus précis, car cela signifie attribuer un sens au résultat produit par le dispositif, sans pour autant comprendre son mode de fonctionnement. En outre, les chercheurs en IHM s'efforcent d'expliquer ce que le modèle peut communiquer à un utilisateur, en privilégiant des approches qui examinent la manière dont l'IA modifie la confiance des utilisateurs vis-à-vis du système (Arrieta *et al.*, 2020 ; Cai *et al.*, 2019 ; Cheng *et al.*, 2019 ; Cramer *et al.*, 2008 ; Shin, 2021 ; Martijn *et al.*, 2022).

Pour les informaticiens, l'explicabilité n'a donc pas de définition stabilisée : pour les chercheurs en ML, expliquer signifie *prédire*, tandis que pour les chercheurs en IHM, expliquer signifie *interpréter*. Dans les deux cas, les solutions proposées demeurent avant tout mathématiques et sont principalement conçues par des informaticiens. C'est pourquoi les chercheurs en SHS soulignent qu'une pleine compréhension du phénomène exige de replacer le dispositif technique dans le contexte de sa conception et de son usage. Ainsi, le débat autour de l'explicabilité est placé par ces derniers dans la myriade d'intérêts économiques et politiques divers, parfois concurrents, qui se manifestent à l'égard de l'IA.

La remise en contexte des outils, des systèmes et des utilisateurs

L'étude de l'IA en sciences humaines et sociales ne se limite pas à l'analyse des algorithmes ou des opérations prévisibles qu'ils exécutent. Elle vise à replacer ces dispositifs techniques dans les contextes sociaux, culturels et matériels où ils prennent sens. Autrement dit, il ne s'agit pas uniquement de comprendre comment l'IA fonctionne, mais de s'interroger sur les usages qu'en font les acteurs, sur les situations concrètes dans lesquelles elle est mobilisée, et sur l'équilibre entre les coûts et les bénéfices attendus pour la collectivité (Beaudouin *et al.*, 2020).

Comme le soulignent Chateauraynaud et Lamy (2025), ces techniques nécessitent d'être situées dans le flux des activités humaines et des environnements matériels dans lesquels elles s'inscrivent. Une telle approche permet de dépasser une lecture technocentrée et solutionniste pour envisager l'IA comme un phénomène sociotechnique, traversé par des logiques d'appropriation, de contestation et de reconfiguration des pratiques. Vuarin et Steyer (2023) explorent, par exemple, les dimensions managériales de l'intégration de l'IA, afin de montrer que l'adoption de ces techniques entraîne une reconfiguration des compétences et de l'organisation du travail, aussi bien pour les métiers existants que pour ceux à venir. À ce titre émergent des professionnels spécialisés, comme les *prompt engineers*, capables de traduire les demandes des utilisateurs en requêtes adaptées, facilitant ainsi l'obtention de réponses plus performantes.

Le principal apport des travaux en sciences humaines et sociales réside donc dans la mise en contexte des pratiques de l'explicabilité, qui dépasse la simple description des relations de cause à effet en intégrant des dimensions socio-économiques plurielles. Notre postulat est que l'étude de cas des industries culturelles offre la possibilité de préciser pour qui, dans

quel contexte et pour quels motifs des formes de transparence peuvent être attendues.

LES USAGES DE L'IA DANS LES INDUSTRIES CULTURELLES

L'IA est devenue omniprésente et généralement utilisée dans des secteurs variés tels que la santé (Mignot, 2025), les transports (Pidoux, Kypraiou et Dehaye, 2025) ou encore les services publics et l'action de l'État (Kirtchik et Musiani, 2025). L'IA fait partie des technologies *enabling* (Brey, 2017) qui renouvellent une large gamme de secteurs et d'industries, en même temps qu'elle se combine facilement avec d'autres techniques pour créer d'autres produits et services. Les industries culturelles n'échappent pas à ce mouvement général, tout en y apportant des spécificités propres à leur secteur. En effet, tout au long de la chaîne de valeur, l'IA ouvre un large éventail d'applications pour les acteurs des industries culturelles. En phase de production, elle permet d'analyser les données de marché afin d'anticiper le succès d'un contenu (Latreille de Fozières, 2025). Du côté de la consommation, elle est mobilisée dans les systèmes de recommandation qui orientent les choix des utilisateurs (Seaver, 2022 ; Farchy *et al.*, 2017). Enfin, en amont, elle s'intègre progressivement aux processus de création (Anichini et Geffroy, 2021). Dans ce contexte, nous proposons d'examiner ces trois champs d'application à travers le prisme de la littérature sur l'XAI.

Prédire et promouvoir les futurs succès

L'économie de la culture étant en partie une économie de prototype, où chaque œuvre est unique et soumise à une forte incertitude quant à sa réception, un petit nombre de productions capte l'essentiel de la demande, selon le modèle de « l'économie des superstars » proposé par Rosen (1981). L'exploitation de données grâce à des algorithmes redonne vigueur à l'ambition de prendre les décisions d'investissement adaptées et d'appuyer voire de remplacer les habituelles intuitions et expertises humaines par des analyses supposées objectives des déterminants du succès d'une œuvre ou d'un artiste. Au-delà du rêve, largement inatteignable, de parfaitement modéliser les clés de la réussite, plus prosaïquement des algorithmes sont déjà largement utilisés pour repérer les tendances du marché, faciliter la prise de décision pour produire un titre ou encore préciser la stratégie commerciale afin de cibler, d'élargir le public potentiel ou d'optimiser la présence d'une œuvre ou d'un artiste selon les supports de diffusion.

Outre l'analyse des tendances du marché, l'une des promesses des IA est de comparer, à partir de l'exploitation de données historiques, les contenus ayant connu le succès avec ceux en cours de production afin d'analyser les clés de la réussite, et éventuellement de l'anticiper. La rupture avec les techniques quantitatives classiques repose sur le fait que la modélisation est issue des données elles-mêmes. Ces traitements, fondés sur l'apprentissage automatique, sont intrinsèquement conservateurs ; ils n'anticipent pas des évolutions, mais reproduisent le passé dans le présent ou l'avenir pour fournir un résultat, des tendances ou estimations. Dans le cadre de cet usage de l'IA, la performance du modèle ne tient pas à l'explicabilité de ses résultats, mais à sa capacité à fournir rapidement et efficacement une réponse au problème posé.

Recommander des contenus en personnalisant les propositions

Le terme de recommandation renvoie à différents dispositifs qui orientent les choix de l'usager et participent à la mise en avant de certains contenus disponibles sur un catalogue numérique. Pour les plateformes et applications en ligne, des formes classiques de recommandation éditorialisées, communes à tous les usagers, restent très présentes notamment sur les services musicaux ou audiovisuels (Farchy *et al.*, 2017). Cependant, les traitements algorithmiques, grâce à l'exploitation automatisée de grandes quantités de données, ont connu un essor considérable et ce sont les recommandations personnalisées issues de ces traitements qui font l'objet de toutes les attentions. La recommandation et la

personnalisation sont devenues les piliers de la stratégie de croissance et d'expérience utilisateur des géants du streaming. Ce modèle correspond au « courtage informationnel » : les plateformes servent d'intermédiaire entre producteurs et publics, tirant profit de cette relation, sans créer de contenus (Miège *et al.*, 2013). Dans ce contexte, le comportement des utilisateurs est encodé et analysé en permanence afin d'affiner les choix des titres susceptibles de les satisfaire par la suite. Cela n'a rien de passif, car les usagers recourent à des dispositifs autonomes d'exploration, tels que les moteurs de recherche et les discographies d'artistes, pour recadrer quand nécessaire le choix qui est fait de manière automatique par les services de streaming musical. La délégation aux algorithmes ne représente qu'un peu moins d'un quart des écoutes (Beuscart, Coavoux et Maillard, 2019). Pour autant, cela signifie-t-il que l'usager s'intéresse au fonctionnement interne de l'algorithme ?

Dans certains contextes, il peut en effet chercher à comprendre pour quelles raisons l'IA a généré un résultat donné, notamment lorsqu'il le perçoit comme insatisfaisant (« pourquoi l'algorithme me propose-t-il un contenu qui ne me correspond pas ? »). En revanche, la transparence sur les mécanismes de recommandation modifie-t-elle leur manière d'interagir avec ces systèmes ? Louafi et M'Barki (2024) apportent des éléments de réponse à cette question. Leur analyse expérimentale repose sur un dispositif de laboratoire dans lequel les participants interagissent avec un système de recommandation semi-personnalisé. Ce cadre expérimental place les utilisateurs dans une situation qui les incite à réfléchir au fonctionnement de l'algorithme et à leurs propres pratiques d'écoute. Les auteurs observent ainsi que l'explication de l'algorithme induit un changement dans les comportements de découverte musicale : plus les individus s'informent sur la recommandation, plus ils ont tendance à écouter les morceaux proposés, renforçant ainsi une posture de « consommation» plutôt que d'« exploration ». Ces résultats encore largement exploratoires laissent néanmoins entendre que les enjeux les plus significatifs de transparence dans les industries culturelles ne se manifestent pas aujourd'hui au niveau de la prédiction des succès ni de la recommandation de contenus aux usagers, mais plutôt en amont, du côté de ce qui alimente le « réacteur » des « industries de l'imaginaire » (Flichy, 1991) : la création.

Transformer la création

Le mouvement d'implication de l'IA dans la création est en effet désormais bien lancé ; deux catégories se distinguent nettement dans notre analyse de la littérature existante.

Dans le premier cas, il s'agit d'accompagner le processus humain de création, de faciliter le travail et l'inspiration du créateur en élargissant son champ des possibles. Les acteurs qui se positionnent ainsi n'ont pas pour but de vendre des productions, mais de proposer de l'accompagnement aux créateurs, en leur fournissant des outils grâce auxquels ils peuvent se libérer de contraintes techniques, de temps ou matérielles. Ils accompagnent les utilisateurs « aguerris » (musiciens, compositeurs, programmeurs) en leur permettant d'intervenir sur un grand nombre de paramètres. L'IA contribue alors à abaisser le seuil de compétence requis, en simplifiant l'accès à certaines fonctionnalités et certains logiciels autrefois réservés à des groupes professionnels bien délimités (cadreurs, photographes, monteurs, créateurs d'effets spéciaux, etc.) dans des domaines d'activité plus classiques (Butraud, Da Silva et Méadel, 2026).

Dans le second cas, au contraire, il s'agit de s'émanciper le plus largement possible du travail du créateur en faisant émerger des œuvres complètement nouvelles qui limitent fortement l'intervention humaine. Il s'agit ici plus spécifiquement de l'irruption de l'IA générative sur le marché grand public, à travers des applications commerciales qui proposent des solutions « clés en main ». Ces outils permettent à des utilisateurs néophytes, soucieux de leur budget ou contraints par le temps, d'obtenir rapidement des compositions prêtes à l'emploi. Les entreprises qui développent ces technologies attirent une clientèle en quête de contenus générés automatiquement : bien que ces productions ne soient pas toujours de haute qualité, elles présentent une valeur commerciale en raison de leur accessibilité, de leur

faible coût et de leur capacité à répondre à une demande de production de masse.

Dans les deux cas, transparence et explicabilité deviennent des enjeux pragmatiques bien loin de principes éthiques généraux. La création, au cœur des industries culturelles, voit son économie bouleversée par l'IA, et l'IAG en particulier sous deux aspects sur lesquels nous mettons l'accent : les conditions de travail (partie 3) et les formes possibles de régulation (partie 4).

PROCESSUS CRÉATIF ET MAÎTRISE DE L'IA

L'intégration de l'IA dans les routines de travail suscite des questionnements dans une large variété de secteurs et d'industries. Dans le cadre médical, par exemple, Anichini et Geffroy (2021) remarquent que l'IA promet une pratique radiologique plus objective en renforçant la précision du diagnostic. Cependant, elle se heurte aux savoirs tacites des radiologues, c'est-à-dire aux normes informelles et implicites qui guident leur interprétation. Ainsi, pour des diagnostics où l'incertitude est faible, l'usage de l'IA peut significativement se substituer au travail des radiologues ; en revanche, dans les cas plus complexes, l'expertise du radiologue demeure indispensable pour interpréter les résultats et établir un diagnostic final, en intégrant des éléments que l'IA ne peut pas appréhender pleinement. Cette complémentarité souligne l'importance d'une collaboration entre outils techniques et savoir-faire, permettant à l'IA d'apporter une plus-value, sans se substituer au travail humain. Dans les industries culturelles, quelles sont les conséquences d'une absence de maîtrise de l'IA sur la création ?

Sur cette question, l'étude de Doshi et Hauser (2024), fondée sur un échantillon de 300 auteurs ayant volontairement recours à l'IA générative pour la rédaction de textes, montre qu'à l'échelle individuelle, cette technologie stimule la créativité des utilisateurs en les exposant à des idées et des références extérieures à leur univers habituel. En revanche, si l'IA favorise la créativité individuelle et améliore la structuration des textes, à l'échelle collective, les récits produits avec son aide tendent à se ressembler davantage. En effet, le modèle génère des réponses standardisées fondées sur un raisonnement statistique, à partir d'une base de données commune, ce qui conduit, malgré la singularité des requêtes, à une homogénéisation des récits. Les résultats et la démarche méthodologique de cette étude peuvent être discutés, notamment leur évaluation très quantitative, nécessairement réductrice, de la créativité des participants. Néanmoins, cet article ouvre la voie à la réflexion sur la délégation de compétences à l'IA. Le risque est que les utilisateurs s'appuient sur ces outils avant d'avoir pleinement développé leurs propres capacités en matière d'écriture. Dans ce contexte, comprendre le fonctionnement de l'IA joue un rôle clé. Cela permet de mieux saisir le rôle de la machine dans le processus créatif et de voir qui malgré les promesses qui sont faites par les géants de la *tech*, ces outils restent limités en matière de création de contenu (Bender et Hanna, 2025). L'explicabilité, située dans le contexte d'usage, favorise une intégration de l'IA dans les industries culturelles, de manière à en tirer parti sans pour autant renoncer à la maîtrise du résultat final.

Corriger, recadrer et intervenir

Les créateurs délèguent certaines tâches à l'IA sans toujours être en mesure d'interpréter les décisions qui ont conduit à un résultat donné. Ce qui entre et sort de la « boîte noire » ne correspond pas toujours aux attentes de l'utilisateur. L'enjeu pour l'utilisateur ne réside pas tant dans l'assurance du bon fonctionnement technique de la machine, qui nécessiterait une maîtrise avancée des codes et des systèmes informatiques, que dans la compréhension des raisons pour lesquelles certains rendus peuvent sembler inintelligibles, erronés ou inattendus. La question abordée ici est celle de la délégation : à quelles conditions peut-on réellement faire confiance à une génération de contenus automatisée ? La répartition des responsabilités entre l'utilisateur et la machine devient alors un enjeu central, à la fois pour éviter les malentendus et pour favoriser des résultats pertinents. Dans ce contexte,

l'explicabilité telle que précédemment définie, joue un rôle fondamental : elle facilite l'intervention humaine en permettant de corriger des erreurs, de reformuler un prompt ou encore de demander des précisions afin d'améliorer la qualité du résultat final.

Les vidéos générées par l'IAG tels que *Luma, Gen3, Kling* et *Sora* ne peuvent généralement être utilisées en l'état, nécessitant donc corrections et améliorations de la part de l'utilisateur. Cette intervention demande de nouvelles compétences et représente un coût économique non négligeable avec le recours à une main-d'œuvre externalisée, fragmentée et précarisée, indispensable au fonctionnement de l'IA (Poiroux, 2023 ; Le Ludec, 2024). En explorant l'intégration des modèles d'IAG dans l'industrie des jeux vidéo, Tailleur et Ramis (2025) contestent l'idée selon laquelle les mondes virtuels n'auront plus besoin de *designers* et de programmeurs dans un avenir proche et soulignent que l'intervention humaine *a posteriori* représente une contrainte significative qui fait appel à des compétences spécifiques pour que les développeurs soient en mesure de maîtriser le processus et d'assumer la responsabilité des résultats produits.

Faire de l'erreur une source de création

En plaçant ce qu'il entend par *erreur* (résultats inadaptés, pas fiables ou incohérents) au centre de la démarche créative, l'utilisateur de l'IA explore de nouvelles formes de créativité. Le duo d'artistes Sofia Crespo et Feileacan McCormick met souvent en lumière les lacunes des bases de données sur lesquelles l'IA s'appuie. En manipulant des millions de données issues de collections d'animaux et de plantes pour recréer des espèces, ils constatent que lorsque les données disponibles sont insuffisantes, l'IA produit des images avec des morphologies et des couleurs détachées de la réalité. Ces « erreurs » deviennent alors le cœur de leur projet artistique, révélant non seulement l'inégalité dans la représentation des espèces sur les bases de données, mais aussi la façon dont les IA sont limitées par les données qui les nourrissent. Cette approche explore la manière dont l'erreur peut entraîner le créateur à penser différemment : parce que l'usager apprend à utiliser les IA par l'expérimentation, en inventant par exemple de nouveaux prompts ou en obtenant des résultats différents pour la même requête, il élargit le champ de ses pratiques par l'exploration. En outre, cela donne la possibilité de dépasser les conceptions traditionnelles de l'erreur, perçue non plus comme une simple déviation à corriger, mais comme une valorisation des accidents, des échecs et des résultats inattendus comme des points de départ pour l'imagination et l'innovation (Auray, 2007 ; Merton et Barber, 2004).

LA TRANSPARENCE POUR RENDRE DES COMPTES ET RÉGULER

Dans la compréhension du fonctionnement général de l'IA, la tension entre des opérateurs de l'IA et les États apparaît de plus en plus palpable, en particulier dans les relations entre États-Unis d'Amérique, Chine et Europe, chacun adoptant une stratégie différente. Les États-Unis soutiennent leurs géants technologiques dans une logique de leadership géopolitique, la Chine mise sur un développement intégré de l'IA, étroitement contrôlé par l'État, tandis que le premier texte de loi d'envergure émane de l'Union européenne. Le règlement européen 2024/1689 du 13 juin 2024 (dit IA act), établissant des règles harmonisées concernant l'IA a en effet alimenté de nombreux débats entre la nécessité de réguler au nom de principes éthiques et la volonté de ne pas freiner l'innovation des entreprises. Pour le régulateur européen, l'explicabilité n'a plus comme objectif prioritaire de comprendre, mais de rendre des comptes (Busuioc, 2021), de pouvoir auditer les modèles d'IA pour faire appliquer les lois existantes et respecter des principes jugés prioritaires.

Si des obligations de transparence pour les fournisseurs et déployeurs de systèmes d'IA irriguent l'ensemble de l'IA Act, les fondements juridiques de ces obligations restent plus flous (Brugière, 2025). Cependant, le législateur exprime des arbitrages politiques. Ainsi, certaines obligations de transparence portent plus particulièrement sur les systèmes considérés comme à « haut risque » (Article 13). De même, la demande de transparence est

explicite en matière d'exploitation des données personnelles ou pour que s'exerce le droit de la concurrence – en particulier en évitant les phénomènes d'auto-préférence des GAFAM, consistant à se nourrir de leurs propres données sans que leurs concurrents n'y aient accès.

Concernant plus spécifiquement les industries culturelles et les médias, on trouve plusieurs interventions autour de la question du droit d'auteur. Il n'en fallait pas plus pour que de nombreux ayant droit se saisissent de la transparence comme du socle d'une demande de respect des droits de propriété intellectuelle et de rémunération lors de l'entrainement des modèles d'IA sur la base de données protégées. Historiquement, en matière de droit d'auteur, la transparence est en effet un sujet ancien et très sensible, que l'IA ne fait que réactualiser.

Deux sujets, évoqués au fil de divers articles du règlement, peuvent être distingués en la matière, celui de la transparence des sources et celui de la transparence des résultats.

Premièrement, en ce qui concerne les modèles d'IA à usage général, dès l'introduction de l'IA Act il est fait mention de la transparence sur les sources :

> « Afin d'accroître la transparence concernant les données utilisées dans le cadre de l'entraînement préalable et de l'entraînement des modèles d'IA à usage général, y compris le texte et les données protégés par la législation sur le droit d'auteur, il convient que les fournisseurs de ces modèles élaborent et mettent à la disposition du public un résumé suffisamment détaillé du contenu utilisé pour entraîner les modèles d'IA à usage général » (p.28).

L'article 53, indiquant les obligations incombant aux fournisseurs de modèles d'IA à usage général, fait spécifiquement le lien avec la question des droits d'auteur. En amont, au niveau des sources, l'IA ne peut fonctionner sans les données massives qui l'alimentent. Les opérateurs d'IA empruntent de manière indifférenciée des données numériques sans toujours citer la source d'œuvres pourtant protégées. Sur le plan économique, l'accès aux données culturelles représente en effet un enjeu essentiel pour assurer le respect des chaînes de valeur. L'aspiration de ces données étant de nature à bousculer les équilibres des secteurs concernés, leur valorisation devra impérativement être assurée dans un souci de soutenabilité des modèles d'affaires des acteurs de l'IA tout en respectant des formes de rémunération susceptibles d'inciter à créer des œuvres nouvelles.

Deuxièmement, en ce qui concerne les résultats générés par des IA, on trouve également des obligations de transparence dans l'article 50 du règlement européen :

> « Lorsque le contenu fait partie d'une œuvre ou d'un programme manifestement artistique, créatif, satirique, fictif ou analogue, les obligations de transparence énoncées au présent paragraphe se limitent à la divulgation de l'existence de tels contenus générés ou manipulés d'une manière appropriée qui n'entrave pas l'affichage ou la jouissance de l'œuvre. Les *déployeurs* d'un système d'IA qui génère ou manipule des textes publiés dans le but d'informer le public sur des questions d'intérêt public indiquent que le texte a été généré ou manipulé par une IA ».

En aval, l'algorithme produit des résultats, fruits d'une collaboration homme-machine et l'on peut donc se demander si cette réalisation finale peut être qualifiée d'œuvre de l'esprit par le droit et, dans cette hypothèse, qui est l'auteur et qui est le titulaire des droits. Parce que la frontière entre création assistée par une IA et création générée par une IA est, dans la pratique, difficile à tracer, le statut juridique de ces créations se révèle complexe et leur identification transparente par les utilisateurs nécessaire afin que puissent être détectés « les deep fakes ».

La distinction effectuée dans l'article 50 entre « deep fakes artistiques » et « deep fakes informationnels » est intéressante dans la mesure où elle souligne que, s'ils incarnent les deux facettes d'une même technologie, ils relèvent de logiques et d'implications distinctes. Tout d'abord, les *deep fakes* informationnels, conçus pour manipuler l'information ou

diffuser des fausses nouvelles, représentent un risque significatif pour le travail journalistique au sein des médias. En matière artistique, les *deep fakes* permettent d'explorer des formes de création, souvent en poussant les limites des effets spéciaux ou en enrichissant les œuvres. On peut alors s'interroger sur la justification d'une exigence de transparence au nom de l'information du public lorsqu'il s'agit d'œuvres revendiquant explicitement le statut de fictions. Le public doit-il être toujours informé des trucages et effets spéciaux qui permettent à la fiction d'exister ? Face à une collaboration étroite entre artistes, techniciens et modèles d'IA, l'exigence de transparence nous semble relever d'une autre logique que celle de l'intégrité du débat public ; il s'agit plutôt d'accompagner les mutations de la chaîne de valeur des productions hybrides qui résultent de la collaboration hommes-machines.

CONCLUSION

Comme nous venons de le voir dans cet article qui se veut méta-théorique, l'injonction à la transparence s'impose comme un impératif largement partagé, mais souvent réduit à une référence abstraite à l'éthique. L'analyse de la littérature en XAI révèle que cette demande repose d'abord sur des dispositifs d'explicabilité à visée technique, que les sciences humaines et sociales invitent à recontextualiser dans des usages situés et traversés par des enjeux socio-économiques et politiques. Les industries culturelles offrent à cet égard un terrain particulièrement riche pour observer ces dynamiques. L'IA y est mobilisée à toutes les étapes de la chaîne de valeur : pour analyser les tendances du marché en amont de la production, pour orienter les choix des utilisateurs via des systèmes de recommandation, et désormais pour participer directement aux processus créatifs eux-mêmes. À mesure que les IA génératives gagnent en accessibilité, la question n'est plus seulement de comprendre de quelle manière les systèmes fonctionnent, mais de savoir comment répartir les responsabilités et garantir la qualité et la diversité de la création. Ainsi, l'explicabilité devient un levier essentiel pour maintenir l'humain dans la boucle, en permettant d'intervenir, de corriger, de contester ou d'ajuster les productions générées. Elle participe à établir une relation de confiance, tout en contribuant à la montée en compétence des utilisateurs et à la reconnaissance de nouveaux savoir-faire. Elle permet également de penser autrement les erreurs ou les biais, en les intégrant comme des ressources potentielles pour la création. L'explicabilité de l'IA apparaît comme un objet de négociation, dont les formes, les finalités et les effets varient selon les contextes, les acteurs et les intérêts en présence. Étudier cette notion à partir du cas des industries culturelles permet d'analyser la régulation dans un écosystème technique en recomposition et de concurrence internationale exacerbée.

NOTES

[1] "Research Priorities for Robust and Beneficial Artificial Intelligence : An Open Letter", Future of Life Institute, publié le 28 octobre 2015, en ligne.

[2] Étienne Caillebotte, « Chiffres ChatGPT : les statistiques à connaître en 2024 », Blog du modérateur, publié le 22 avril 2024, en ligne.

[3] Cette recherche bénéficie d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-23-PEIC-0006 (PEPR ICCARE/STYX).

RÉFÉRENCES BIBLIOGRAPHIQUES

Adadi, Amina ; Berrada, Mohammed (2018), « Peeking Inside the Black-Box : A Survey on Explainable Artificial Intelligence (XAI) », *IEEE Access*, n° 6, p. 52138-60, doi : 10.1109/ACCESS.2018.2870052.

Afchar, Darius ; Melchiorre, Alessandro ; Schedl, Markus ; Hennequin, Romain ; Epure, Elena, et Moussallam ; Manuel (2022), « Explainability in Music Recommender Systems », *AI Magazine*, vol. 43, n° 2, p. 190-208, doi : 10.1002/aaai.12056.

Afchar, Darius (2023), « Interpretable Music Recommender Systems », *Thèse de doctorat*, Sorbonne Université.

Anichini, Giulia ; Geffroy, Bénédicte (2021), « L'intelligence artificielle à l'épreuve des savoirs tacites. Analyse des pratiques d'utilisation d'un outil d'aide à la détection en radiologie », *Sciences sociales et santé*, vol. 39, n° 2, p. 43-69, doi : 10.1684/sss.2021.0200.

Arrieta, Alejandro Barredo ; Díaz-Rodríguez,Natalia ; Del Ser, Javier ; Bennetot, Adrien ; Tabik, Siham ; Barbado, Alberto ; García, Salvador ; Gil-López, Sergio, Daniel ; Molina, Richard Benjamins ; Chatila, Raja ; Herrera, Francisco (2019), « Explainable Artificial Intelligence (XAI) : Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI », *Information Fusion*, vol. 58, n°1, p. 82–115, doi :10.1016/j.inffus.2019.12.012.

Attencourt, Boris ; Berrebi-Hoffmann, Isabelle ; Lamy, Jérôme ; Tighanimine, Mariame (2025), « Le travail des algorithmes : sociologie des délégations techniques », *Socio*, vol. 20, p. 7-16.

Auray, Nicolas (2007), « Folksonomy : A New Way to Serendipity », *Communications & Stratégies*, vol. 65, n° 1, p. 67-91.

Barnes, Susan B. (2006), « A Privacy Paradox : Social Networking in the United States », *First Monday*, vol. 11, n° 9, doi : 10.5210/fm.v11i9.1394.

Beaudouin, Valérie ; Bloch, Isabelle ; Bounie, David ; Clémençon, Stéphan ; Florence, d'Alché-Buc ; Eagan, James ; Maxwell, Winston ; Mozharovskyi, Pavlo ; Parekh, Jayneel (2020), « Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach », *SSRN Electronic Journal*, 2020, pp.1-66.

Beaudouin, Valérie ; Velkovska, Julia (2023), « Enquêter sur l'"éthique de l'IA" », *Réseaux*, vol. 240, n° 4, p. 9-27, doi : 10.3917/res.240.0009.

Bélisle-Pipon, Jean-Christophe ; Monteferrante, Erica ; Roy, Marie-Christine ; Couture, Vincent (2023), « Artificial Intelligence Ethics Has a Black Box Problem », *AI & Society*, vol. 38, n° 4, p. 1507-1522, doi : 10.1007/s00146-021-01380-0.

Bender, Emily M. ; Hanna, Alex (2025). *The AI Con : How to Fight Big Tech's Hype and Create the Future We Want*. London : The Bodley Head.

Beuscart, Jean-Samuel ; Coavoux, Samuel ; Maillard, Sisley (2019), « Les algorithmes de recommandation musicale et l'autonomie de l'auditeur : Analyse des écoutes d'un panel d'utilisateurs de streaming », *Réseaux*, vol. 213, n° 1, p. 17-47, doi : 10.3917/res.213.0017.

Brey, Philip A. E. (2017), « Ethics of Emerging Technologies » (p. 175-192), in Hansso, Sven Ove (dir.), *The Ethics of Technology : Methods and Approaches*, Lanham : Rowman & Littlefield.

Brugière, Jean-Michel (2025), « IA générative, transparence partout, justification nulle part!», *Recueil Dalloz*, vol. 3, p. 120.

Butraud, Anouck ; da Silva, Jaércio ; Méadel, Cécile (2026), « Framing AI in the audiovisual industries on LinkedIn » in Micalizzi, Alessandra (dir.), *Artificial Intelligence and Social Research: Methods, Contexts, Imaginaries*, Rome : WriteUp Books, p. 175-200.

Busuioc, Madalina (2021), « Accountable Artificial Intelligence : Holding Algorithms to Account », *Public Administration Review*, vol. 81, n° 5, p. 825-36, doi : 10.1111/puar.13293.

Cai, Carrie J. ; Jongejan, Jonas ; Holbrook, Jess (2019), « The effects of example-based explanations in a machine learning interface » (p. 258-62) in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, New York : Association for Computing Machinery.

Callon, Michel (2006), « Sociologie de l'acteur-réseau » (p. 267-76), in Akrich, Madeleine ; Latour, Bruno (dir.), *Sociologie de la traduction : Textes fondateurs*, Paris : Presses des Mines.

Callon, Michel, Akrich, Madeleine ; Dubuisson-Quellier, Sophie (2013), *Sociologie des agencements marchands : textes choisis*, Paris : Presses de Mines.

Callon, Michel ; Méadel, Cécile ; Rabeharisoa, Vololona (2000), « L'économie des qualités », *Politix*, vol. 4, n° 52, p.211-239, doi : 10.3406/polix.2000.1126.

Chateauraynaud, Francis (2019), « Petit traité de contre-intelligence artificielle. Retour sociologique sur des expérimentations numériques », *Zilsel*, vol. 5, n° 1, p. 174-95, doi : 10.3917/zil.005.0174.

Chateauraynaud, Francis ; Lamy, Jérôme (2025), « Les algorithmes et leurs écologies », *Socio*, vol. 20, p. 17-40.

Cheng, Hao-Fei, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper et Haiyi Zhu (2019), « Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders » (p. 1-12) in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow: ACM.

Cramer, Henriette ; Evers, Vanessa ; Ramlal, Satyan ; van Someren, Maarten ; Rutledge, Lloyd ; Stash, Natalia ; Aroyo, Lora ; Wielinga, Bob (2008), « The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender », *User Modeling and User-Adapted Interaction*, vol. 18, n° 5, p. 455-496, doi : 10.1007/s11257-008-9051-3.

Crawford, Kate (2021), Atlas of AI : *Power, politics, and the planetary costs of artificial intelligence*, New Haven : Yale University Press.

Crépel, Maxime ; Cardon, Dominique (2022), « Robots vs algorithmes : Prophétie et critique dans la représentation médiatique des controverses de l'IA », *Réseaux*, vol. 232-233, n° 2, p. 129-67, doi : 10.3917/res.232.0129.

Doshi, Anil R. et Oliver P. Hauser (2024), « Generative AI enhances individual creativity but reduces the collective diversity of novel content », *Science Advances*, vol. 10, n° 28, p. 5290, doi : 10.1126/sciadv.adn5290.

Doshi-Velez, Finale ; Kim, Been (2017), « Towards A Rigorous Science of Interpretable Machine Learning ».

Farchy, Joëlle ; Anciaux ; Arnaud ; Méadel, Cécile (2017), « Une question de comportement. Recommandation des contenus audiovisuels et transformations numériques », *Tic & Société*, vol. 10, n° 2-3, p. 168, doi : 10.4000/ticetsociete.2136.

Flichy, Patrice (1991), *Les Industries de l'imaginaire : pour une analyse économique des médias*, Grenoble : Presses Universitaires de Grenoble.

Guidotti, Riccardo ; Anna, Monreale ; Ruggieri, Salvatore ; Turini, Franco ; Pedreschi, Dino; Giannotti, Fosca (2018), « A Survey Of Methods For Explaining Black Box Models », *ACM Comput. Surveys*, vol. 51, n°5, doi : 10.1145/3236009.

Gunning, David (2017), « Explainable artificial intelligence (XAI), 2017 », *DARPA/I2O*.

Jobin, Anna ; Ienca, Marcello ; Vayena, Effy (2019), « The Global Landscape of AI Ethics

Guidelines », *Nature Machine Intelligence*, vol. 1, n° 9, p. 389-99, doi : 10.1038/s42256-019-0088-2.

Kirtchik, Olessia (2019), « STS et Intelligence artificielle : une rencontre manquée ? », *Zilsel*, vol. 5, n° 1, p. 149-60, doi : 10.3917/zil.005.0149.

Kirtchik, Olessia ; Musiani, Francesca (2025), « Vers un « État-automate » ? », *Socio*, vol. 20, p. 101-126.

Latreille de Fozières, Noé (2025), « La réflexivité algorithmique », *Socio*, vol. 20, p. 101-126.

Le Ludec, Clément (2024), « Des humains derrière l'intelligence artificielle. La sous-traitance du travail de la donnée entre la France et Madagascar », Thèse de sociologie, Institut Polytechnique de Paris.

Lipton, Zachary C. (2018), « The Mythos of Model Interpretability : In machine learning, the concept of interpretability is both important and slippery », *Queue*, vol. 16, n° 3, p. 31–57, doi : 10.1145/3236386.3241340.

Louafi, Mehdi ; M'Barki, Julien (2024), « Algo-Rhythm Unplugged : Effects of Explaining Algorithmic Recommendations on Music Discovery », *SSRN*, http://dx.doi.org/10.2139/ssrn.4982393

Martijn, Millecamp ; Conati, Cristina ; Verbert, Katrien (2022), « "Knowing me, knowing you": Personalized explanations for a music recommender system », *User Modeling and User-Adapted Interaction*, vol. 32, n° 1-2, p. 215-52, doi : 10.1007/s11257-021-09304-9.

Merton, Robert K. ; Barber, Elinor (2004), *The Travels and Adventures of Serendipity : A Study in Sociological Semantics and the Sociology of Science*, Princeton : Princeton University Press.

Miège, Bernard ; Bouquillion, Philippe ; Mœglin, Pierre (2013) « L'industrialisation des biens symboliques Les industries créatives en regard des industries culturelles ». Fontaine : Presses universitaires de Grenoble. « Communication médias, société ».

Mignot, Léo (2025), « Déléguer aux algorithmes ? », *Socio*, vol. 20, p. 83-100.

Miller, Tim (2019). « Explanation in artificial intelligence : Insights from the social sciences», *Artificial Intelligence*, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Pidoux, Jessica ; Kypraiou, Sofia ; Dehaye, Paul-Olivier (2025), « Gaining transparency in Uber's algorithmic management », *Socio*, vol. 20, p. 41-64.

Poiroux, Jérôme (2023), « La fabrique des algorithmes : conception et impact au sein des organisations. Une sociologie des processus d'engagement et de désengagement en régime computationnel » Thèse de sociologie, EHESS.

Van de Poel, Ibo (2017), « Society as a Laboratory to Experiment with New Technologies » (p. 404) in Bowman, D. M. ; Stokes E. ; A. Rip, Jenny (dir.), *Embedding New Technologies into Society : A Regulatory, Ethical and Societal Perspective*, Stanford in the Vale : Stanford Publishing.

Ribeiro, Marco Tulio ; Singh, Sameer ; Guestrin, Carlos (2016), « "Why Should I Trust You ? »: Explaining the Predictions of Any Classifier ». *arXiv*, Conference Paper, https://doi.org/10.48550/arXiv.1602.04938.

Rosen, Sherwin (1981), « The Economics of Superstars », *The American Economic Review*, vol. 71, n° 5, p. 845-858.

Rudin, Cynthia (2019), « Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead », *Nature Machine Intelligence*, vol. 1, p. 206–215.

Seaver, Nick (2022), *Computing Taste : Algorithms and the Makers of Music Recommendation*, Chicago : University of Chicago Press.

Shin, Donghee (2021), « The effects of explainability and causability on perception, trust, and acceptance : Implications for explainable AI », *International Journal of Human-Computer Studies*, vol. 146, p. 1025-1051, doi : 10.1016/j.ijhcs.2020.102551.

Tailleur, Gabriel ; Ramis, Morgane (2025), « Metaverse and AI diffusion : An empirical assessment from a financial perspective », *IEEE Engineering Management Review*, in Press, 53 (3), pp.1-13.

Van Lent, Michael ; Fisher, William ; Mancuso, Michael (2004), « An explainable artificial intelligence system for small-unit tactical behavior ». *In Proceedings of the 16th conference on Innovative applications of artifical intelligence (IAAI'04)*. AAAI Press, 900–907.

Vuarin, Louis ; Steyer, Véronique (2023), « Le principe d'explicabilité de l'IA et son application dans les organisations », *Réseaux : communication, technologie, société*, vol. 240, n° 4, p. 179-210, doi : 10.3917/res.240.0179.

Weitz, Katharina ; Schiller, Dominik ; Schlagowski, Ruben ; Huber, Tobias ; André, Elisabeth (2021), « "Let Me Explain!" : Exploring the Potential of Virtual Agents in Explainable AI Interaction Design », *Journal on Multimodal User Interfaces*, vol. 15, n° 2, p. 87–98, doi : 10.1007/s12193-020-00332-0.