****

**L’EXPLICABILITÉ DANS L’IA**

**Recommandations politiques pour les législateurs américains**

**Janine ECKER,**

**Paul KLEINEIDAM, Claudia LEOPARDI,**

**Anna PADIASEK, Benjamin SALDICH**

**Etudiants**

**Master en politiques publiques à l’Ecole d’affaires publiques de Sciences Po.**

**Spécialité Digital, New Technology and Public Policy**

**Mai 2024**

**January 2024**

**Table des matières**

Résumé…………………………………………………………………………….…..3

Introduction……………………………………………………………………………4

Chapitre 1: **Pourquoi les régulations actuelles ne marchent pas ?** 6

*1. Le régime publicitaire……………………………………………………… 8*

*2. L'introduction de l'IA générative………………………………………. 9*

Chapitre 2: Que valent les tentatives régulationnistes de reprise en main de la Commission Européenne ? 10

2.1. Dénoncer les méchants et les mettre hors de capacité de nuire …………10

2.2. Renforcer la rapidité de réaction des organismes de contrôle ……………12

2.3. Faire pression sur les plateformes pour adopter des bonnes pratiques… 13

2.4. Eduquer les utilisateurs………………………………………………….. 15

# **Chapitre 3: Les plateformes de réseaux sociaux ont engendré un milieu de propagations à haute fréquence…………………………………….18**

3.1. Les biais conceptuels des intentions régulationnistes actuelles……………18

* *Contenus déjà faits ou propagations en train de se faire…….. .19*
* *Free speech ou free reach…………………………………………20*
* *La focalisation sur les émetteurs…………………………………..21*
* *S'attaquer à des contenus spécifiques……………………………22*
* *Ralentir seulement les mauvaises propagations…………………22*

3.2. Une analyse épidémiologie de la viralité des réseaux sociaux ……………23

 *D: La durée………………………………………………………… 23*

*O: Les opportunités……………………………………………….. 25*

*T: La transmissibilité………………………………………………...28*

*S: La susceptibilité………………………………………………… 30*

Conclusion 32

Liste de recommandations 34

**Bibliographie……………………………………………………………………… 36**

1. Résumé

À une époque où l'intelligence artificielle (IA) est appelée à jouer un rôle de plus en plus important dans notre société, il est impératif de maintenir un certain niveau de contrôle humain sur les systèmes d'IA. L'explicabilité - définie de manière générale comme un niveau de compréhension ou d'explication du fonctionnement et des décisions des systèmes d'IA - est une composante essentielle de ce contrôle humain. Pourtant, les universitaires, les éthiciens et les législateurs ne sont pas parvenus jusqu'à présent à se rallier à une stratégie unique pour réglementer l'explicabilité dans le domaine de l'IA. Cette note politique, produite par notre groupe de réflexion européen, synthétise les connaissances universitaires et les approches réglementaires internationales afin de proposer des recommandations applicables aux décideurs politiques américains. Notre objectif est de trouver un équilibre entre les impératifs éthiques et les considérations pratiques, en garantissant la transparence, la responsabilité et la confiance de la société dans les technologies de l'IA.

Après avoir examiné la compréhension actuelle des notions de transparence dans les systèmes d'IA "boîte blanche" et "boîte noire", le document analyse la manière dont les organisations et les pays ont cherché à définir et à réglementer l'explicabilité de l'IA, en se concentrant plus particulièrement sur l'UE, la Chine et les États-Unis. Cette analyse fait ressortir trois stratégies politiques principales, dont les forces et les limites sont examinées.

S'inspirant des récents efforts de réglementation de l'UE, ce document recommande une approche équilibrée de l'explicabilité de l'IA qui vise à réglementer la gouvernance de l'IA en fonction des niveaux de risque, en reconnaissant les limites techniques tout en garantissant la responsabilité et la transparence. Nous proposons quatre stratégies politiques clés que le Congrès américain devrait prendre en compte lors de l'élaboration de la législation sur l'IA :

1. Mettre en œuvre une approche fondée sur le risque : L'adoption d'un cadre structuré semblable à la loi européenne sur l'IA garantit la cohérence, la transparence et la proportionnalité de la réglementation en matière d'IA.
2. Imposer des obligations contraignantes aux systèmes à haut risque : Renforcer la transparence et les approches centrées sur l'humain pour les systèmes d'IA à haut risque, afin de garantir la responsabilité et d'atténuer les risques.
3. Établir des règles claires en matière de responsabilité : introduire des règles de responsabilité pour faciliter la réparation des préjudices causés par les systèmes d'IA, en équilibrant les mesures préventives et les mécanismes de réparation des préjudices.
4. Création d'un groupe de travail de la FTC : créer un groupe de travail spécialisé au sein de la FTC pour superviser la gouvernance de l'IA, garantir la conformité et encourager la collaboration entre les parties prenantes.

Ce document souligne également la complexité et la nature évolutive du secteur de l'IA, qui pose des défis uniques à la conception et à la mise en œuvre d'une réglementation centrée sur l'explicabilité. L'obtention d'explications fiables pour la prise de décision en matière d'IA reste un défi important et doit être abordé dans le cadre de recherches futures.

**Partie 1 : Introduction**

"Je suis surpris qu'il soit si bas", a déclaré James Rivelli, lorsqu'on lui a annoncé que son évaluation du risque COMPAS n'était que de 3 sur 10.[[1]](#footnote-1) Le COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), un algorithme censé déterminer le risque de récidive des prévenus, a été utilisé dans tout le système de justice pénale des États-Unis, aidant les juges à déterminer les peines pénales en attribuant un score de risque à chaque prévenu. Rivelli avait été arrêté pour vol et, malgré un casier judiciaire contenant des agressions graves, des vols et des trafics de stupéfiants, le COMPAS lui a attribué un score de 3, ce qui indique un faible risque de récidive. En comparaison, Robert Cannon, un autre homme arrêté pour vol à l'étalage et dont le casier judiciaire était nettement moins chargé, s'est vu attribuer un risque moyen de 6. La réalité ultérieure de ces deux personnes remettrait en question l'évaluation du risque effectuée par l'algorithme : James Rivelli a été condamné à deux chefs d'accusation pour un cambriolage supplémentaire.

La disparité dans l'évaluation des risques entre des individus comme Cannon et Rivelli au sein du système COMPAS souligne l'importance cruciale de l'explicabilité dans les systèmes d'IA et le besoin urgent de cadres juridiques et réglementaires clairs régissant leur développement et leur déploiement. Cela soulève une question fondamentale : Comment pouvons-nous comprendre, justifier ou expliquer les processus de prise de décision algorithmique ? Les journalistes d'investigation de Propublica suggèrent que cette différence d'évaluation pourrait être liée à la race des accusés. Leur analyse a révélé que les Noirs sont presque deux fois plus susceptibles que les Blancs d'être étiquetés comme présentant un risque élevé sans pour autant récidiver, tandis que les Blancs sont plus susceptibles d'être étiquetés comme présentant un risque faible mais de commettre des délits.

Les disparités raciales dans l'évaluation des risques au sein de l'algorithme COMPAS ont probablement eu un impact sur d'innombrables personnes dans le système de justice pénale américain. Toutefois, la méthodologie et la structure qui sous-tendent l'algorithme COMPAS sont protégées en tant que secret commercial et, même si elles étaient rendues publiques, l'explicabilité de l'algorithme reste floue, tant pour ses résultats que pour son processus.

Le COMPAS n'est qu'un des innombrables algorithmes qui, malgré le manque de transparence de leurs processus internes, ont un impact significatif sur les libertés civiles des citoyens américains. En tant que pays, les États-Unis sont à l'origine de nombreuses innovations dans le domaine de l'IA, mais le gouvernement fédéral n'a pas encore réussi à faire en sorte que cette industrie naissante dispose des garde-fous nécessaires pour protéger ses citoyens contre les risques les plus pressants de l'IA, notamment le risque que les humains perdent la capacité de comprendre et de garder le contrôle des systèmes d'IA avec lesquels ils sont de plus en plus souvent en contact chaque jour.

Notre objectif est de contribuer au débat sur la justification, la faisabilité et l'opportunité de l'explicabilité à l'ère de l'IA. En s'appuyant sur les connaissances universitaires et les enseignements tirés des approches réglementaires internationales de la gouvernance de l'IA, cette note politique vise à fournir des recommandations applicables aux décideurs politiques américains. Il propose une voie concrète pour aller de l'avant, en plaidant pour une approche de l'explicabilité qui concilie les impératifs éthiques et les considérations pratiques. En faisant le lien entre les débats théoriques et les applications du monde réel, il vise à contribuer à l'élaboration de réglementations robustes en matière d'éthique de l'IA qui soutiennent la transparence, la responsabilité et la confiance de la société dans les technologies de l'IA. Pour ce faire, nous passerons en revue la littérature existante sur l'explicabilité, nous examinerons les options politiques déjà mises en œuvre dans d'autres pays et, enfin, nous formulerons des recommandations à l'intention des décideurs politiques actuels et futurs, alors que nous entrons rapidement dans une société irrémédiablement façonnée par l'IA.

Dans la section suivante, nous commencerons par définir le concept d'explicabilité et sa signification sémantique.

### L'effet boîte noire : Pourquoi nous avons besoin d'explications

La transparence est un aspect fondamental des systèmes d'intelligence artificielle (IA) éthiques, essentiel pour favoriser la confiance, la responsabilité et l'utilisation responsable de ces technologies.[[2]](#footnote-2) En théorie, comprendre comment les systèmes d'IA parviennent à leurs décisions permettrait de se prémunir contre les biais et de valider les résultats obtenus par l'IA, mais la capacité actuelle à "expliquer" l'IA dépend du système, une distinction méthodologique commune étant faite entre les systèmes d'IA "à boîte blanche" et "à boîte noire".

Dans les systèmes d'IA à boîte blanche, le processus de prise de décision est totalement transparent. Grâce à leur arbre de décision structurel, ces modèles donnent un aperçu clair des mécanismes internes, du traitement des données d'entrée, de la prise en compte des caractéristiques et des étapes de la prise de décision. En revanche, les systèmes d'IA à boîte noire offrent une interprétabilité minimale. Les mécanismes internes opaques de ces systèmes obscurcissent les processus décisionnels, empêchant les utilisateurs de comprendre comment les données d'entrée se traduisent en données de sortie.[[3]](#footnote-3) L'opacité des systèmes d'IA "boîte noire" pose plusieurs problèmes notables, notamment l'impossibilité d'interpréter les résultats, la crainte d'une partialité et d'une discrimination algorithmiques, et l'entrave au respect de la réglementation et à la responsabilité.[[4]](#footnote-4) Toutefois, l'explicabilité des systèmes d'IA peut constituer un moyen prometteur de remédier à l'effet "boîte noire".

### L'explicabilité en tant que principe éthique dans les réglementations juridiques

Le concept d'explicabilité en tant que principe éthique pour l'IA a existé en tant qu'aspect fondamental de la vie privée numérique, avec une version des exigences d'explicabilité incluse dans la loi française sur la protection des données de 1978.[[5]](#footnote-5) Des travaux plus récents rédigés par le Dr Luciano Floridi et approuvés par les lignes directrices éthiques de la Commission européenne sur l'IA digne de confiance (European Commission AI HLEG, 2018) ont exposé ces principes à l'ère actuelle de l'IA. Selon Floridi et al, l'explicabilité est essentielle pour permettre l'application d'autres principes éthiques tels que la bienfaisance, la non-malfaisance, l'autonomie et la justice, et peut être définie comme englobant la transparence, la compréhensibilité, l'interprétabilité et la responsabilité.[[6]](#footnote-6) Les chercheurs affirment que l'explicabilité devrait garantir que les systèmes d'IA sont intelligibles, c'est-à-dire qu'il devrait être possible de répondre aux questions sur leur fonctionnement. Ils reconnaissent toutefois que l'étendue des explications requises peut varier en fonction du contexte du système d'IA et du destinataire de l'explication. Un exemple utilisé par les auteurs indique qu'en médecine, il n'est pas toujours possible de fournir des explications entièrement mécanistes et que les praticiens et les patients peuvent s'appuyer sur des preuves corrélatives. En outre, la profondeur de l'explication nécessaire peut varier en fonction du destinataire, par exemple un médecin par rapport à un patient.[[7]](#footnote-7)

D'un point de vue éthique, l'explicabilité est considérée comme un moyen de responsabilisation. Floridi et al. suggèrent que la responsabilité englobe non seulement les développeurs, mais aussi les commanditaires, les déployeurs et les utilisateurs au sein du système sociotechnique. Cette perspective plus large de la responsabilité permet de reconnaître les efforts déployés pour aider les utilisateurs moins "avertis sur le plan technique" à utiliser l'IA de manière responsable.[[8]](#footnote-8) Toutefois, les auteurs cités ne plaident pas en faveur d'une explicabilité totale des systèmes d'IA dans leurs moindres détails, mais soulignent plutôt l'importance de fournir le niveau d'explication adéquat pour que chaque utilisateur ou partie concernée assume la responsabilité des effets de l'utilisation de l'IA.

### Explicabilité ou explicabilité

La sémantique des termes "explicabilité" et "expliquabilité" fait l'objet d'un débat scientifique important, ce qui conduit à une ambiguïté de ces notions. L'explicabilité fait référence à la capacité d'un système d'IA à clarifier son processus de prise de décision d'une manière compréhensible pour les humains. L'explicabilité, quant à elle, va plus loin. Elle consiste à rendre les systèmes d'IA transparents et compréhensibles pour les parties prenantes, y compris les autorités juridiques, les parties concernées et la société dans son ensemble. L'explicabilité va au-delà de la simple explication et englobe la notion plus large de rendre les systèmes d'IA transparents, compréhensibles et interprétables pour les parties prenantes, y compris les autorités légales, les parties concernées et la société dans son ensemble.[[9]](#footnote-9) Alors que les exigences réglementaires mettent souvent l'accent sur la transparence et l'interprétabilité, l'explicabilité ne se limite pas à ces aspects, en particulier dans les scénarios du monde réel, avec des contraintes de temps et des niveaux variables de compétence technologique.[[10]](#footnote-10) Il est important de préciser que, dans le présent document, les deux termes sont utilisés de manière interchangeable.

### Énoncé du problème

Comme le soulignent Floridi et al, qu'il s'agisse de "transparence", de "responsabilité" ou d'"intelligibilité", chacun de ces principes rend compte d'un aspect apparemment nouveau de l'IA : le fait que ses rouages sont souvent invisibles ou inintelligibles. Cet article se réfère à ces notions distinctes mais étroitement liées sous le nom d'"explicabilité", abordant à la fois l'aspect épistémologique de l'intelligibilité (qui permet de comprendre "comment cela fonctionne-t-il ?") et la dimension éthique de la responsabilité ("qui est responsable de la manière dont cela fonctionne ?").[[11]](#footnote-11) Nous plaidons pour un cadre d'explicabilité qui harmonise les impératifs éthiques avec les considérations pratiques. Notre objectif est de contribuer à l'élaboration de réglementations solides en matière d'éthique de l'IA à l'échelle mondiale, en donnant la priorité à la transparence, à la responsabilité et à la confiance de la société. En intégrant des idées théoriques à des applications du monde réel, nos recommandations visent à fournir des idées applicables aux décideurs politiques.

Dans la section suivante, nous passerons en revue les approches réglementaires actuelles au niveau mondial, ce qui nous aidera à évaluer les forces et les faiblesses des options politiques disponibles aux États-Unis et à formuler des recommandations à l'intention des décideurs politiques. Nous avons choisi de nous adresser spécifiquement aux législateurs américains pour plusieurs raisons. Les politiques adoptées par les États-Unis peuvent avoir d'importantes répercussions dans le monde entier, en influençant les normes et les pratiques mondiales. En conseillant les États-Unis, nous pouvons avoir un impact sur la gouvernance de l'IA dans le pays et à plus grande échelle. En outre, les États-Unis doivent trouver un équilibre entre la nécessité de favoriser l'innovation et l'impératif de protéger les intérêts publics. Il est essentiel d'instaurer et de maintenir la confiance du public dans les technologies de l'IA. Les conseils en matière d'explicabilité - axés sur la transparence, la responsabilité et l'intelligibilité - peuvent contribuer à créer des cadres qui améliorent la compréhension et la confiance du public dans les systèmes d'IA.

# **Partie 2 : Évaluation des options politiques**

Alors que l'influence de l'IA ne cesse de s'étendre, garantir la transparence et l'explicabilité des systèmes d'IA est devenu une priorité cruciale pour les décideurs politiques du monde entier. Cela a conduit à l'élaboration de diverses initiatives politiques dans différentes juridictions visant à réglementer l'IA et à promouvoir des pratiques éthiques. Qu'il s'agisse de mesures législatives ou de lignes directrices pour l'industrie, ces initiatives reflètent un engagement commun en faveur de la responsabilité et de la surveillance de l'IA. Toutefois, la diversité des approches met en évidence la complexité de la question de la transparence et de l'explicabilité de l'IA à l'échelle mondiale.

Dans ce qui suit, nous examinerons diverses initiatives politiques mises en œuvre dans différentes juridictions afin de conceptualiser et d'appliquer l'explicabilité de l'IA. Nous nous concentrerons sur les principales initiatives, notamment celles de l'OCDE, de la Chine et de l'UE. Ces juridictions représentent des intérêts géopolitiques et économiques divers, chacune offrant des perspectives et des approches uniques. Nous avons choisi d'analyser ces initiatives en raison de leur influence substantielle et de leurs contributions significatives à l'élaboration d'une gouvernance mondiale de l'IA. En outre, nous anticipons leur impact réglementaire sur les États-Unis. Leurs actions ne façonnent pas seulement la trajectoire du développement de l'IA dans leurs régions respectives, mais ont également de profondes implications pour la communauté internationale au sens large, par exemple par le biais de mécanismes tels que "l'effet Bruxelles", qui a des effets de facto et de jure sur les États-Unis. [[12]](#footnote-12)

## Initiatives et normes internationales

L'émergence de divers nouveaux documents et cadres politiques à l'échelle mondiale témoigne d'une reconnaissance croissante de l'importance de comprendre la logique qui sous-tend les résultats de l'IA en tant qu'aspect essentiel de la garantie de la fiabilité. Les principes de l'OCDE en matière d'IA, adoptés en mai 2019 et mis à jour en mai 2024 par les 38 pays membres de l'OCDE, constituent un cadre international préliminaire pour la réglementation de l'IA.[[13]](#footnote-13) En tant qu'instrument juridique de l'OCDE, les principes représentent une aspiration commune pour ses pays membres, constituant un ensemble de normes qui visent à promouvoir une utilisation innovante et digne de confiance de l'IA dans le respect des droits de l'homme et des valeurs démocratiques.[[14]](#footnote-14) L'OCDE a publié cinq principes directeurs, dont l'un est la "transparence et l'explicabilité".[[15]](#footnote-15) Il exige des acteurs de l'IA qu'ils "s'engagent à la transparence et à la divulgation responsable concernant les systèmes d'IA et qu'ils fournissent des informations significatives, adaptées au contexte et conformes à l'état de l'art".[[16]](#footnote-16) Dans l'ensemble, cela vise à garantir que les gens comprennent quand ils s'engagent avec des systèmes d'IA et qu'ils sont en mesure de contester les résultats de la prise de décision de l'IA. En outre, l'organisation précise que la compréhension des systèmes d'IA doit reposer sur "une information facile à comprendre sur les facteurs et la logique qui ont servi de base à la prédiction, à la recommandation ou à la décision", invoquant ainsi une norme d'intelligibilité. Ces principes ont ensuite servi de base aux principes du G20 en matière d'IA.[[17]](#footnote-17) Bien qu'ils ne soient pas juridiquement contraignants, les principes de l'OCDE existant dans d'autres domaines politiques ont démontré une influence significative en établissant des lignes directrices internationales et en aidant les gouvernements à élaborer une législation nationale. [[18]](#footnote-18)

Dans ce contexte, un autre aspect essentiel est l'élaboration de normes d'explicabilité. Les normes internationales facilitent l'harmonisation et l'interopérabilité entre les juridictions, ce qui permet une collaboration efficace et l'échange de bonnes pratiques en matière de gouvernance de l'IA à l'échelle mondiale. Elles sont donc essentielles pour garantir l'uniformité et la cohérence de la réglementation, en particulier dans le domaine de l'explicabilité de l'IA, afin de promouvoir l'interopérabilité et de faciliter la collaboration à l'échelle mondiale.

Actuellement, diverses organisations internationales, coalitions et comités façonnent le cadre réglementaire des normes d'explicabilité. Actuellement, la Commission européenne (CE) collabore avec des organismes européens de normalisation (OEN) tels que le CEN, le CENELEC et l'ETSI pour créer des normes à l'appui de la loi européenne sur l'IA.

La CE a publié un projet de demande de normalisation au CEN-CENELEC, soulignant les exigences des normes pour soutenir la présomption de conformité.[[19]](#footnote-19) En outre, l'attention est portée sur le Comité technique mixte 21 "Intelligence artificielle" (JTC 21) du CEN-CENELEC, qui inclut l'explicabilité parmi ses principaux thèmes de recherche. L'Institut européen des normes de télécommunication (ETSI) contribue également à conseiller la loi sur l'IA, en mettant l'accent sur la transparence et l'explicabilité.[[20]](#footnote-20) Aux États-Unis, le National Institute of Standards and Technology (NIST) a produit des documents d'orientation sur les principes d'explicabilité (développés plus loin dans le présent document), tandis que la Federal Trade Commission (FTC) s'intéresse aux pratiques irrégulières en matière d'IA. [[21]](#footnote-21)

Au niveau mondial, les activités de normalisation internationale sont menées par des organismes tels que l'Organisation internationale de normalisation (ISO), la Commission électrotechnique internationale (CEI) et l'Institut des ingénieurs électriciens et électroniciens (IEEE). Par exemple, l'ISO/IEC JTC 1/SC 42 "Intelligence artificielle" se concentre sur la normalisation des programmes d'IA, avec des documents tels que l'ISO/IEC TR 24028:2020 qui considère l'explicabilité comme une mesure d'atténuation des vulnérabilités et des menaces de l'IA.[[22]](#footnote-22) Les efforts de normalisation de l'IEEE comprennent des documents traitant de la partialité algorithmique, de la transparence et des exigences pour que les systèmes d'IA soient reconnus comme explicables.[[23]](#footnote-23) Ces normes internationales sont essentielles pour façonner le paysage de la réglementation de l'IA, faciliter l'interopérabilité et promouvoir un développement responsable de l'IA dans le monde entier.

## Dispositions relatives à l'explicabilité dans les règlements de l'UE

L'approche de l'UE en matière de gouvernance de l'IA se caractérise par l'accent mis sur les valeurs et les principes centrés sur l'homme, en axant les stratégies de gouvernance sur la protection des droits fondamentaux et les considérations éthiques. Le cadre réglementaire de l'UE en matière de technologie donne la priorité à la responsabilité, à la transparence et à l'équité, dans le but d'atténuer les dommages potentiels et de favoriser l'innovation et la compétitivité.

### GDPR

Un premier débat notable concernant les explications relatives aux systèmes de prise de décision algorithmique, y compris l'IA, a émergé en 2018 avec l'établissement du principe du "droit à l'explication" au sein du Règlement général sur la protection des données (RGPD) de l'UE. La mise en œuvre du GDPR marque une étape importante pour répondre aux préoccupations entourant l'opacité de la prise de décision individuelle automatisée (ADM). Le GDPR contient des dispositions concernant la prise de décision automatisée basée sur des données personnelles qui visent à établir des garanties pour les individus qui peuvent être soumis à des décisions uniquement basées sur le traitement automatisé de leurs données personnelles, y compris le profilage (article 22). En particulier, le considérant 71 du GDPR souligne l'importance de garantir des protections aux personnes concernées, y compris le droit d'obtenir une explication de la décision.[[24]](#footnote-24) En outre, l'article 13, paragraphe 2, point f), l'article 14, paragraphe 2, point g), et l'article 15, paragraphe 1, point h), prévoient que les personnes concernées reçoivent des informations utiles sur la logique qui sous-tend l'ADM, ainsi que sur les conséquences potentielles pour l'individu.

L'inclusion de ces dispositions dans le GDPR a suscité des débats juridiques concernant l'établissement d'un "droit à l'explication".[[25]](#footnote-25) Si les partisans de ces dispositions soutiennent qu'elles fournissent un cadre permettant de garantir l'explicabilité des décisions automatisées, leurs détracteurs s'inquiètent de leur imprécision et des difficultés de mise en œuvre pratique qu'elles posent[[26]](#footnote-26) . Il convient de noter que la formulation utilisée dans ces articles du GDPR peut limiter le champ d'application, l'exigence d'"informations significatives" pouvant n'englober que la structure générale et la fonctionnalité d'un système ADM, et non les circonstances individuelles d'une décision spécifique, ce qui limite le champ d'application de l'article. [[27]](#footnote-27)

### Loi européenne sur l'IA

Conformément à sa stratégie réglementaire globale axée sur les valeurs, l'UE a adopté un cadre juridique régissant le développement, le déploiement et l'utilisation de l'IA conformément à ses valeurs fondamentales, comme en témoignent la sécurité et la fiabilité des systèmes d'IA.[[28]](#footnote-28) L'UE a été la première juridiction à proposer et à adopter une législation pour faire face aux risques liés à l'IA par le biais de la loi européenne sur l'IA. Ce règlement contraignant constitue un cadre réglementaire complet régissant le développement, le déploiement et l'utilisation des systèmes d'IA dans les 27 États membres. Il adopte une approche fondée sur les risques, classant les systèmes d'IA en fonction de leur impact potentiel. Les applications à haut risque, telles que celles qui concernent les infrastructures critiques ou les soins de santé, sont soumises à des réglementations plus strictes afin de garantir la sécurité, la transparence et la responsabilité. Certaines applications d'IA menaçant les droits des citoyens sont interdites. Les systèmes à risque moyen ou faible sont soumis à des mesures proportionnées. Ces mesures visent à trouver un équilibre entre les intérêts souvent contradictoires de la promotion de l'innovation et de la protection des droits fondamentaux, en veillant à ce que les technologies de l'IA soient développées et utilisées de manière responsable au sein de l'UE.

La loi sur l'IA contient une terminologie couramment utilisée dans la littérature scientifique sur l'IA explicable (XAI), notamment la transparence, l'opacité et la compréhensibilité.[[29]](#footnote-29) Elle impose la conception et le développement de systèmes d'IA qui présentent une "transparence suffisante", permettant aux utilisateurs "d'interpréter avec précision leurs résultats". [[30]](#footnote-30)

Par conséquent, certaines dispositions de la loi sur l'IA ont été interprétées par certains auteurs comme suggérant des éléments d'explicabilité de l'IA et l'utilisation de méthodologies XAI et de modèles d'IA basés sur la transparence. [[31]](#footnote-31)

La loi sur l'IA s'attaque à l'opacité de manière globale en s'appuyant sur deux piliers, notamment (1) les obligations de transparence et les exigences en matière de surveillance humaine, ainsi que (2) les systèmes de gestion des risques. La loi sur l'IA énonce des exigences de transparence pour les systèmes d'IA à haut risque, visant à garantir que leur fonctionnement est compréhensible et interprétable pour les utilisateurs. L'article 13 de la loi sur l'IA souligne le besoin de transparence en stipulant que ces systèmes doivent être conçus de manière à permettre aux utilisateurs d'interpréter et d'utiliser efficacement leurs résultats. Cette exigence de transparence est renforcée par le considérant 47 de la loi sur l'IA, qui souligne l'importance de la transparence pour remédier à l'opacité et à la complexité potentielles des systèmes d'IA. La loi précise que les systèmes d'IA doivent être accompagnés d'un mode d'emploi clair et complet, fournissant aux utilisateurs des informations pertinentes sur les caractéristiques, les capacités et les limites du système.[[32]](#footnote-32) Toutefois, plutôt que d'imposer des modèles spécifiques de conception transparente ou des outils d'IAO, l'accent est mis sur la possibilité pour les utilisateurs d'utiliser le système de manière appropriée. L'article 14 de la loi sur l'IA met également l'accent sur la surveillance humaine, en veillant à ce que les personnes chargées de superviser les systèmes d'IA possèdent les compétences et l'autorité nécessaires pour contrôler efficacement leur fonctionnement. Le considérant 48 de la loi sur l'IA souligne l'importance des mesures de contrôle humain pour garantir la réactivité des opérateurs humains et se prémunir contre les biais ou erreurs potentiels dans les résultats des systèmes d'IA.

Le système de gestion des risques de la loi sur l'IA, décrit à l'article 9, renforce encore l'importance de l'explicabilité dans l'IA. La reconnaissance sous-jacente est que, malgré les obligations introduites pour les systèmes à haut risque, il se peut que ces systèmes n'atténuent pas suffisamment tous les risques, laissant certains risques non résolus. C'est pourquoi l'article 9 de la loi sur l'IA introduit la nécessité de systèmes de gestion des risques pour garantir que les fournisseurs identifient et évaluent tous les risques restants.[[33]](#footnote-33) Cela souligne l'importance de l'explicabilité pour garantir que les fournisseurs de services d'IA puissent gérer et atténuer efficacement les risques potentiels associés à leurs systèmes.

Bien que la loi européenne sur l'IA englobe une approche globale de la réglementation des systèmes d'IA, elle n'impose pas explicitement l'utilisation de techniques d'IAO ou de modèles transparents de par leur conception. La loi met plutôt l'accent sur la transparence, la surveillance humaine et les mesures de gestion des risques pour remédier à l'opacité des systèmes d'IA. Cette approche permet aux fournisseurs et aux utilisateurs de développer leurs propres méthodes de conformité, compte tenu des limites techniques et des complexités associées aux méthodes d'IAO.

### Proposition de directive de l'UE sur la responsabilité en matière d'IA

La Commission européenne a introduit la directive sur la responsabilité en matière d'IA (AILD) en 2021, parallèlement à la loi sur l'IA.[[34]](#footnote-34) Cette directive vise à faciliter les réclamations civiles pour les dommages subis par les utilisateurs finaux de systèmes d'IA en établissant des règles claires en matière de preuve et de causalité. La directive met en avant la responsabilité civile pour les dommages causés par les systèmes d'IA ou leur incapacité à atteindre les résultats escomptés, en abordant les droits d'indemnisation et autres devoirs de diligence liés aux réglementations sur l'IA.

L'AILD a proposé l'introduction de la "présomption de causalité", qui allège la charge de la preuve pour les victimes en permettant aux tribunaux de présumer que le non-respect des obligations pertinentes a causé un dommage si un lien de causalité avec les performances de l'IA est raisonnablement probable, bien que cette présomption puisse être réfutée. L'article 4, paragraphe 2, point b), rendrait ainsi obligatoire l'obligation d'explication pour les systèmes d'IA complexes, opaques et à haut risque qui ne satisfont pas aux exigences de transparence énoncées à l'article 13 de la loi sur l'IA. En outre, il permet aux victimes d'accéder aux preuves nécessaires en demandant la divulgation d'informations concernant les systèmes d'IA à haut risque, ce qui contribue à identifier les parties responsables et à comprendre la cause du préjudice, tout en garantissant la protection des données sensibles telles que les secrets commerciaux. [[35]](#footnote-35)

L'AILD n'est encore qu'une proposition en attente d'examen par le Parlement européen et le Conseil de l'Union européenne. Néanmoins, elle témoigne de l'engagement de l'UE à garantir la transparence et la responsabilité des systèmes d'IA, en particulier dans le domaine de l'explicabilité. En introduisant des dispositions relatives à l'établissement de la responsabilité civile et des demandes d'indemnisation pour les dommages résultant du non-respect des réglementations en matière d'IA, la directive AILD démontre l'approche proactive de l'UE pour relever les défis associés aux systèmes d'IA opaques. [[36]](#footnote-36)

## Dispositions relatives à l'explicabilité dans la réglementation chinoise

Ces dernières années, le gouvernement chinois a commencé à introduire des réglementations complexes et multiples dans la sphère numérique. À certains égards, comme la protection des données personnelles, la Cyberspace Administration of China, l'autorité de régulation numérique du pays, a calqué ses politiques sur celles de ses homologues européens. Dans d'autres domaines, la Chine semble être en tête de la course à la réglementation, comme dans le cas de la réglementation des algorithmes ou des lignes directrices pour le développement de l'IA.[[37]](#footnote-37) Les efforts du pays pour introduire des règles dans les nouvelles technologies comprennent également des efforts pour rendre cette technologie compréhensible pour ses utilisateurs et ceux qui sont affectés par ses conséquences.

Les dispositions relatives à l'explicabilité, qui s'appliquent directement ou indirectement aux systèmes d'IA, figurent à la fois dans la législation officielle chinoise et dans les lignes directrices non contraignantes de l'État.[[38]](#footnote-38) Les dispositions législatives relatives à l'explicabilité figurent dans la loi chinoise sur la protection des informations personnelles, le règlement sur les recommandations relatives aux algorithmes, le règlement sur la synthèse profonde ou le règlement sur l'IA générative introduit par la CAC. De nombreux documents d'orientation ont également été publiés par le ministère chinois de la science et de la technologie, tels que les normes éthiques pour l'IA de nouvelle génération publiées en 2021. [[39]](#footnote-39)

Les dispositions légales relatives à l'explicabilité dans le droit chinois font référence au droit des personnes à comprendre les décisions automatisées et au droit des utilisateurs à recevoir des explications lorsqu'un algorithme a un impact majeur sur leurs intérêts.[[40]](#footnote-40) Faisant écho à l'article 22 du GDPR, l'article 24 de la loi de 2021 sur la protection des informations personnelles stipule explicitement qu'une personne a le droit de demander une explication au responsable du traitement de ses informations si les données ont fait l'objet d'une prise de décision automatisée ayant un impact significatif sur ses droits et ses intérêts.[[41]](#footnote-41) En outre, le même droit d'explication a été incorporé dans le règlement sur les recommandations d'algorithmes de 2022. La loi accorde aux utilisateurs individuels le droit de désactiver les recommandations algorithmiques, mais aussi de demander et de recevoir une explication de la part du fournisseur d'un algorithme sur la manière dont le service influe sur les décisions de l'utilisateur.[[42]](#footnote-42) Le règlement prévoit que les services susceptibles d'influencer l'opinion publique doivent fournir aux utilisateurs des informations de base sur les mécanismes algorithmiques et procéder à des auto-évaluations de la sécurité des algorithmes afin de garantir le respect des normes.[[43]](#footnote-43) En outre, en vertu du règlement sur l'IA générative, les fournisseurs de solutions d'IA ont l'obligation, à la demande des autorités de régulation, d'expliquer les mécanismes algorithmiques ainsi que la source, l'échelle et le type de données d'entraînement utilisées par leurs modèles.[[44]](#footnote-44) Ainsi, l'État chinois a commencé à introduire des mesures d'explicabilité à la fois au niveau national, en institutionnalisant les droits des citoyens à recevoir une explication, et dans le cadre de la coopération entre l'État et les entreprises en matière d'algorithmes d'IA.

## Efforts d'explicabilité du statu quo aux États-Unis

En l'absence de législation spécifique à l'IA, l'exploration la plus complète de l'explicabilité des systèmes d'IA au sein du gouvernement fédéral des États-Unis provient du National Institute for Standards and Technology (NIST), un organisme qui fait partie du ministère du commerce et qui se concentre sur les industries technologiques. Les "Quatre principes de l'intelligence artificielle explicable" du NIST formulent une conception américaine de l'explicabilité, axée à la fois sur les "résultats" et les "processus", qui englobe non seulement le raisonnement qui sous-tend une décision particulière prise par un système d'IA, mais aussi l'architecture, la conception et la structure sous-jacentes du système d'IA en question.[[45]](#footnote-45) Le NIST propose des principes autour desquels l'explicabilité peut être conçue : l'explication en soi est une composante nécessaire des systèmes d'IA (1), cette explication doit être significative pour le consommateur visé (2) et précise (3), et les systèmes doivent fonctionner dans les limites de leurs connaissances (4).[[46]](#footnote-46) Ces principes postulent que l'explicabilité augmentera la confiance dans les systèmes d'IA, ce qui aura des effets positifs en aval sur l'adoption par les utilisateurs et l'efficacité, et atténuera les risques d'un système à boîte noire. Le NIST a produit d'autres documents sur l'explicabilité, dont le plus récent est le cadre de gestion des risques liés à l'intelligence artificielle (Artificial Intelligence Risk Management Framework - AI RMF), une ressource disponible en 2023 pour les ministères fédéraux, les organisations et les autres acteurs de l'IA afin de mieux comprendre et d'atténuer les risques associés aux technologies de l'IA, qui s'aligne sur les "quatre principes" du NIST.[[47]](#footnote-47) Toutefois, ces recommandations n'ont aucune autorité juridique, ce qui limite intrinsèquement la portée des normes établies par l'institut.

Les propositions visant à réglementer l'IA, telles que l'*Algorithmic Accountability Act de* 2022, n'ont jusqu'à présent pas été adoptées par le Congrès. Cette proposition aurait exigé que les systèmes d'IA réglementés évaluent la "transparence et l'explicabilité" de leurs systèmes, y compris les "facteurs pertinents qui contribuent à une décision particulière", mais n'aurait imposé aucune stipulation aux entreprises au-delà de l'évaluation et de l'établissement de rapports.[[48]](#footnote-48) Au lieu de cela, la première tentative réussie de réglementation de l'IA aux États-Unis a eu lieu le 30 octobre 2023, lorsque le président Biden a signé un décret décrivant les objectifs de l'approche de son administration en matière de réglementation de l'intelligence artificielle.[[49]](#footnote-49) Bien que la notion de libertés civiles et de droit individuel à la vie privée soit bien établie dans la jurisprudence américaine, les États-Unis n'ont pas adopté de législation complète sur la protection de la vie privée numérique et ne peuvent donc pas s'appuyer sur les protections existantes de la vie privée numérique pour garantir la transparence et l'explicabilité des systèmes d'IA. Il faut donc s'écarter considérablement de l'approche de l'Union européenne en matière de réglementation de l'IA, qui fonctionne théoriquement en tandem avec le GDPR pour garantir le droit des individus à la transparence. Au lieu de cela, le décret cherche à habiliter les secrétaires de divers départements à adapter la législation préexistante à l'ère de l'IA, comme la législation sur la protection de la vie privée de 2002 et le Fair Credit Reporting Act de 1970. [[50]](#footnote-50)

Le décret maintient une notion relativement vague de l'explicabilité, ne mentionnant que brièvement la transparence et la responsabilité. Dans la section 6, le décret exige du secrétaire au travail qu'il "élabore et publie des principes et des bonnes pratiques à l'intention des employeurs qui pourraient être utilisés pour atténuer les inconvénients potentiels de l'IA pour le bien-être des employés et maximiser ses avantages potentiels", y compris "les implications pour les travailleurs de la collecte et de l'utilisation par les employeurs de données les concernant liées à l'IA, notamment la transparence, l'engagement, la gestion et les activités protégées par les lois sur la protection des travailleurs".[[51]](#footnote-51) Plus loin, dans la section 8, il encourage les agences de régulation indépendantes à "utiliser tout l'éventail de leurs pouvoirs pour [...] souligner ou clarifier les exigences et les attentes liées à la transparence des modèles d'IA et à la capacité des entités réglementées à expliquer leur utilisation des modèles d'IA".[[52]](#footnote-52) Ces deux brèves références à l'explicabilité constituent l'intégralité de ce qui pourrait être considéré comme un mandat d'explicabilité, bien qu'aucune ne contienne de termes suffisamment forts pour être considérée comme une application.

Sans le solide mécanisme d'application de la législation fédérale, le décret du président Biden se lit comme une liste de mandats pour la recherche future, obligeant les agences du gouvernement fédéral à analyser les ramifications de l'intelligence artificielle dans leurs domaines de réglementation. En outre, ce décret souligne le manque de mécanismes d'application actuellement à la disposition de M. Biden et démontre le besoin urgent d'un ensemble de réglementations numériques globales aux États-Unis, dont une partie serait axée sur l'explicabilité des systèmes d'intelligence artificielle.

Dans l'ensemble, l'approche du gouvernement fédéral des États-Unis en matière d'explicabilité est vaste, mais sa portée est vague et elle n'est que théorique dans le cadre réglementaire actuel. Cela dit, ces documents témoignent d'un désir d'intégrer une certaine version de l'explicabilité dans la législation sur l'IA, si une telle réglementation devait être introduite dans les années à venir. Bien que les contours exacts de cette approche ne soient pas encore formulés, il semblerait qu'elle fasse la distinction entre les explications relatives au processus et au résultat, et qu'elle mette l'accent sur la signification pour le destinataire, qu'il s'agisse d'un utilisateur, d'un régulateur ou d'un spécialiste de l'industrie.

##

## Options politiques

L'analyse des cadres réglementaires de l'IA, en particulier dans l'UE, aux États-Unis, en Chine et dans le reste du monde, montre que les décideurs politiques ont adopté diverses stratégies pour l'explicabilité de l'IA. L'examen de cet ensemble diversifié d'initiatives internationales fait ressortir trois approches politiques :

1. L'approche du laissez-faire, qui suppose qu'aucune réglementation significative des XAI n'est nécessaire ;
2. Une approche d'interdiction, imposant des obligations d'explicabilité pour toutes les technologies d'IA, interdisant effectivement la fourniture des algorithmes qui ne peuvent être expliqués à leurs utilisateurs ;
3. Une approche équilibrée, composée de lignes directrices et d'obligations plus souples pour les fournisseurs de services d'IA, en fonction de l'impact qu'un algorithme peut avoir sur les individus ou la société en général.

*Figure 1. Approches possibles de la réglementation des XAI*



### L'approche du laissez-faire

L'approche du laissez-faire permet le développement sans restriction de nouvelles technologies, ce qui est important pour garantir la position des entreprises américaines d'IA sur le marché. Avec l'introduction de réglementations, les fournisseurs d'IA pourraient décider de déplacer leur développement hors des États-Unis, ce qui aurait un impact sur la croissance de l'économie américaine. En outre, des études telles que Mohammadi et al (2020) ont montré qu'il existe un compromis entre la maximisation du bien-être total (et de l'utilité du consommateur) et l'obligation pour les entreprises de fournir des explications complètes sur l'IA.[[53]](#footnote-53) Lorsque les fournisseurs d'IA sont tenus d'introduire la XAI dans leurs services, ils doivent non seulement supporter des coûts de mise en œuvre et s'exposer à des violations de la propriété intellectuelle, mais ils ne sont pas non plus en mesure de fournir des explications complètes sur leurs algorithmes.[[54]](#footnote-54) Au lieu de cela, lorsque le marché de l'IA n'est pas réglementé et que la fourniture de XAI reste donc facultative, les entreprises les plus développées se tournent vers la fourniture de XAI à leurs utilisateurs, afin de faire de l'explicabilité un aspect différenciateur de leur produit.[[55]](#footnote-55) Dans le cadre d'un laissez-faire, les instruments politiques concrets sont les suivants :

* Lignes directrices volontaires en matière de transparence : Encourager les développeurs d'IA et les entreprises à adopter des lignes directrices volontaires en matière de transparence
* Auto-réglementation du secteur : Favoriser les initiatives prises par l'industrie pour établir des organismes d'autorégulation ou des normes industrielles visant à promouvoir l'explicabilité.
* Exigences en matière de divulgation publique : Encourager les développeurs d'IA à divulguer publiquement des informations sur leurs systèmes d'IA, y compris les sources de données, les méthodes d'entraînement et les biais potentiels, afin d'améliorer la transparence et d'instaurer la confiance avec les utilisateurs.

### L'approche de l'interdiction

En revanche, l'approche stricte préconise une réglementation rigoureuse, imposant l'explicabilité pour toutes les technologies d'IA, ou pour celles qui dépassent un certain seuil de risque. Cette approche impose des obligations strictes aux développeurs d'IA pour qu'ils fournissent des explications claires et interprétables sur leurs algorithmes, que ce soit sous la forme de voies de décision *a priori* ou d'explications *a posteriori.*[[56]](#footnote-56) Plus précisément, les systèmes d'IA qui ne peuvent pas fournir d'explications à leurs utilisateurs sont effectivement interdits de déploiement. Les modèles simples et auto-interprétables ("boîtes blanches") pourraient être obligés de fournir aux utilisateurs le chemin de décision d'un modèle pour produire un certain résultat, publiant ainsi essentiellement le mode de fonctionnement d'un algorithme donné.[[57]](#footnote-57) Pour les réseaux neuronaux plus complexes, il conviendrait plutôt d'utiliser des explications *a posteriori*. Les explications *post hoc* sont générées après la création du modèle et peuvent être divisées en explications globales ou locales.[[58]](#footnote-58) Obliger les développeurs d'IA à fournir des explications globales d'un système d'IA signifie fournir aux individus une compréhension globale d'un modèle en termes de processus de prise de décision (par exemple, en montrant quels types de données sont pris en compte par un algorithme pour prendre une décision). Inversement, la fourniture d'explications locales peut également être exigée, concernant des résultats spécifiques, afin de clarifier le comportement d'un modèle dans un cas particulier. Ces explications sur le fonctionnement du modèle peuvent être introduites à l'aide de systèmes décomposables ou de modèles de substitution qui se rapprochent des processus décisionnels des boîtes noires.[[59]](#footnote-59) Dans le cadre d'une approche d'interdiction, les options politiques concrètes sont les suivantes :

* Explications obligatoires : Adopter une législation exigeant que tous les systèmes d'IA, en particulier ceux utilisés dans des applications à haut risque, fournissent des explications claires et interprétables sur leurs décisions et leurs actions.
* Interdiction des systèmes d'IA opaques : Interdire le déploiement de systèmes d'IA opaques qui ne peuvent pas fournir d'explications sur leurs résultats, ce qui revient à interdire l'utilisation d'algorithmes "boîte noire" dans des domaines critiques.
* Supervision réglementaire : Mettre en place des organismes de réglementation chargés de contrôler et de faire respecter les exigences en matière d'explicabilité, de procéder à des audits et d'imposer des sanctions en cas de non-respect.

### Approche équilibrée

L'approche équilibrée cherche à trouver un juste milieu entre le laisser-faire et des réglementations plus strictes, en adoptant une approche nuancée et sensible au contexte de la gouvernance de l'IA. Dans le cadre de cette approche, les décideurs politiques reconnaissent les limites techniques et pratiques de l'explicabilité totale et choisissent donc d'imposer des lignes directrices et des obligations aux développeurs d'IA, le plus souvent en fonction des risques potentiels qui doivent être évalués avant la mise sur le marché du système. Les systèmes d'IA à haut risque, tels que ceux utilisés dans les infrastructures critiques ou les soins de santé, sont soumis à des réglementations plus strictes, notamment en matière de transparence et de surveillance humaine. À l'inverse, les systèmes d'IA à faible risque peuvent être régis par des lignes directrices moins prescriptives, ce qui permet une plus grande flexibilité et une plus grande innovation. Notre analyse de la réglementation internationale montre que les approches gouvernementales actuelles, telles que celles de l'UE et de la Chine, entrent dans cette catégorie. Dans le cadre de cette approche intermédiaire, les instruments politiques concrets comprennent :

* Réglementation fondée sur les risques : Mettre en œuvre une approche de la réglementation de l'IA fondée sur les risques, en classant les systèmes d'IA dans différents niveaux de risque en fonction de leur impact potentiel sur les individus et la société.
* Exigences d'explicabilité adaptées : Adapter les exigences réglementaires et les obligations en fonction du niveau de risque des systèmes d'IA, les systèmes à haut risque étant soumis à des exigences d'explicabilité plus strictes et les systèmes à faible risque à des lignes directrices plus souples.
* Bacs à sable réglementaires : Créer des bacs à sable réglementaires ou des environnements expérimentaux où les développeurs d'IA peuvent tester et innover avec de nouvelles technologies sous contrôle réglementaire, ce qui permet de développer des bonnes pratiques et des normes.

###

### Défis et limites réglementaires

Alors que les décideurs politiques expriment largement leur volonté de garantir l'explicabilité de l'IA, le principal défi ne réside pas seulement dans la création de cadres juridiques, mais aussi dans les complexités techniques de la mise en œuvre. Actuellement, les développeurs d'IA sont confrontés à des obstacles considérables pour intégrer l'explicabilité dans leurs systèmes, en grande partie à cause de définitions floues et de l'absence d'approches normalisées. Même l'élaboration hypothétique de normes internationales en matière d'explicabilité est entravée par diverses difficultés et tensions, ce qui complique la faisabilité de tels efforts.

À bien des égards, l'explicabilité des systèmes d'IA complexes semble en contradiction avec la nature même de la technologie. La conception complexe des réseaux neuronaux profonds et d'autres modèles d'IA avancés leur permet de générer des réponses et des solutions nuancées, bien au-delà des capacités des systèmes simplistes basés sur des arbres de décision. Toutefois, ce succès rend également ces systèmes susceptibles de rencontrer des difficultés semblables à celles que rencontrent les humains lorsqu'ils justifient leurs décisions ou leurs actions. Les humains ont souvent du mal à fournir un raisonnement hiérarchique pour justifier leurs décisions, ce qui met en évidence la difficulté inhérente à l'explicabilité dans tout système complexe de prise de décision.[[60]](#footnote-60)

Les réseaux neuronaux et d'autres modèles d'IA complexes sont confrontés aux mêmes problèmes et, par conséquent, les experts en IA n'ont pas encore trouvé de solutions significatives aux mesures explicatives telles que la robustesse (définie comme le fait que l'explication soit "cohérente et précise pour une gamme d'entrées"), la fidélité (le fait que l'explication capture avec précision le processus sous-jacent) et la compréhensibilité (définie comme la capacité des êtres humains à comprendre l'explication donnée).[[61]](#footnote-61) Tout simplement, il n'existe actuellement aucune approche permettant d'obtenir des explications fiables et solides pour la prise de décision de l'IA, du moins pour les modèles qui requièrent le plus d'explications.

Cela pourrait justifier une approche réglementaire plus stricte, dans laquelle les systèmes d'IA inexplicables seraient purement et simplement interdits au-delà d'un certain seuil de risque pour l'homme. Toutefois, les systèmes d'IA à haut risque ont déjà été intégrés dans presque tous les secteurs de notre économie, ce qui rend une interdiction générale des systèmes d'IA inexplicables à haut risque très risquée. L'interdiction de ces systèmes nécessiterait probablement le retrait de composants technologiques clés d'une multitude de produits existants, y compris, mais sans s'y limiter, les fonctions d'assistance à la conduite et de reconnaissance des objets dans les automobiles, l'IA augmentée dans les appareils chirurgicaux ou les systèmes de détection des crises cardiaques dans les opérateurs d'appels d'urgence. Cela aurait un impact immédiat et important sur l'économie nationale, affaiblirait le secteur américain de l'IA et entraverait gravement la capacité des entreprises américaines à rester des leaders de l'industrie et, surtout, créerait en soi des risques importants pour les humains dans l'ensemble de la société. Bien sûr, la législation pourrait prévoir des exceptions pour garantir que ces produits qui sauvent des vies restent sur le marché, mais de telles exceptions remettent en question l'objectif même d'un mandat d'explicabilité.

À l'inverse, nous soutenons qu'une approche de laisser-faire comporte des risques importants en matière d'explicabilité, dont beaucoup ont déjà été évoqués tout au long de ce document. Ces risques comprennent des biais dans les ensembles de données d'apprentissage qui affectent les processus et les résultats des systèmes d'IA et conduisent à la discrimination, à la violation des droits d'auteur et de la propriété intellectuelle, ainsi qu'à d'importantes violations de la vie privée en raison de la nature inexplicable des résultats et des décisions de l'IA. Ces risques ne sont pas seulement théoriques ; il s'agit de menaces tangibles et existantes pour la société ; ce sont les risques auxquels les États-Unis sont actuellement confrontés en l'absence de mesures réglementaires. Négliger de les traiter pourrait exacerber les disparités raciales et de genre et alimenter à juste titre la méfiance à l'égard des systèmes d'IA, en particulier parmi les communautés marginalisées.

Sur la base de ces considérations, nous suggérons qu'une approche réglementaire équilibrée sera essentielle pour naviguer entre les risques concurrents et les limites associées à la fois à une réglementation excessive et à une réglementation insuffisante. Pour parvenir à cet équilibre, il faudra examiner attentivement divers facteurs, notamment les incidences potentielles sur l'innovation, l'équité, la responsabilité et la confiance de la société dans les technologies de l'IA.

# **Partie 3 : Recommandations juridiques**

La dernière partie de notre document présente des recommandations concernant l'explicabilité de l'IA dans le contexte américain. Tirant les leçons de multiples contextes réglementaires nationaux, ces suggestions visent à fournir des protections complètes aux utilisateurs de systèmes d'IA aux États-Unis, tout en continuant à permettre au secteur de l'IA en plein essor de prospérer. Nos recommandations s'articulent autour de quatre grandes actions :

|  |
| --- |
| 1. Mettre en œuvre une approche fondée sur les risques
2. Imposer des obligations contraignantes pour les systèmes à haut risque en adoptant une approche centrée sur l'homme
3. Établir des règles claires en matière de responsabilité afin de faciliter la réparation pour les personnes lésées par les systèmes d'IA
4. Création d'un groupe de travail de la Commission fédérale du commerce (FTC) chargé de superviser la mise en œuvre de l'explicabilité de l'IA
 |

1. **Mettre en œuvre une approche fondée sur les risques**

Nous encourageons vivement les États-Unis à adopter une approche fédérale de la réglementation de l'IA fondée sur les risques, à l'instar du cadre de la loi européenne sur l'IA. La tendance actuelle à un patchwork ascendant de décrets et de réglementations étatiques entrave considérablement la capacité à mettre en place une réglementation efficace, complète et coordonnée de l'ensemble du paysage de l'IA.[[62]](#footnote-62) Au contraire, l'adoption d'une approche fondée sur les risques fournit un cadre descendant qui garantit la cohérence, la transparence et la proportionnalité de la réglementation de l'IA, ce qui est essentiel pour encourager l'innovation, protéger les valeurs sociétales et maintenir la compétitivité mondiale.

Une approche de l'évaluation des systèmes d'IA fondée sur le risque, en particulier en ce qui concerne l'explicabilité, offre plusieurs avantages notables. Tout d'abord, en classant les applications d'IA en fonction de leurs risques potentiels, les régulateurs peuvent orienter leurs efforts de surveillance de manière plus efficace, en concentrant les ressources là où elles sont le plus nécessaires. Les systèmes à haut risque, qui peuvent impliquer des prises de décision complexes ou présenter un potentiel de préjudice plus important, font l'objet d'un examen plus approfondi en ce qui concerne l'explicabilité, ce qui permet de garantir que les mesures de transparence sont correctement hiérarchisées. Deuxièmement, cette approche permet d'établir des exigences et des normes personnalisées pour les systèmes d'IA en fonction de leur profil de risque. Par exemple, les domaines critiques tels que les soins de santé ou la finance peuvent nécessiter des mesures d'explicabilité plus robustes pour garantir la sécurité et la fiabilité. Troisièmement, le fait de fournir des lignes directrices et des attentes claires en matière d'explicabilité permet aux développeurs de faire des choix de conception proactifs, en donnant la priorité à la transparence et à l'interprétabilité de leurs systèmes d'IA. Cela favorise une culture de développement responsable de l'IA et contribue à la création de systèmes intrinsèquement explicables.

En outre, la flexibilité inhérente à une approche fondée sur le risque encourage l'innovation en permettant un environnement réglementaire plus permissif pour les applications à faible risque. Cela favorise l'expérimentation et l'exploration de nouvelles approches de l'explicabilité de l'IA sans trop contraindre les développeurs avec des charges réglementaires, et favorise une industrie plus compétitive. Enfin, une approche fondée sur le risque facilite l'amélioration continue des pratiques de gouvernance de l'IA au fil du temps. À mesure que la technologie évolue et que de nouveaux risques apparaissent, les régulateurs peuvent adapter leurs exigences et leurs normes afin de garantir que les systèmes d'IA restent transparents, responsables et conformes aux valeurs sociétales. Ce processus itératif d'amélioration favorise l'apprentissage et le progrès continus en matière de gouvernance de l'IA. Dans l'ensemble, une approche fondée sur le risque fournit un cadre équilibré pour promouvoir la transparence, la responsabilité et la fiabilité des systèmes d'IA, en encourageant l'innovation tout en maintenant l'adaptabilité de la réglementation.

1. **Imposer des obligations contraignantes pour les systèmes à haut risque en adoptant une approche centrée sur l'homme**

Nous plaidons pour la mise en œuvre d'obligations contraignantes, en particulier pour les systèmes à haut risque (HRS), en mettant l'accent sur une approche centrée sur l'homme pour les protections réglementaires. Bien que nous considérions que l'interdiction pure et simple de ces systèmes soit irréalisable, il est crucial d'atténuer les risques qui leur sont associés par des mesures juridiques. Reconnaissant que l'explicabilité dans l'IA est encore en évolution et n'est pas universellement applicable, il est crucial d'appliquer des obligations qui garantissent une transparence maximale, en particulier dans les HRS. Cela inclut des obligations de transparence strictes qui doivent être imposées aux développeurs et aux déployeurs de HRS. Ces obligations devraient inclure une documentation complète détaillant la conception du système, les données d'entraînement, les algorithmes et les processus de prise de décision. En outre, il devrait y avoir des mandats pour des audits et des évaluations réguliers de ces systèmes afin de garantir la conformité avec les normes de transparence. En outre, des mécanismes devraient être mis en place pour faciliter l'examen externe et la validation des systèmes d'IA par des experts indépendants et des autorités de régulation.

En outre, outre la transparence, il est essentiel de se prémunir contre la dépendance à l'égard de systèmes d'IA inexplicables pour la prise de décision dans des scénarios à haut risque. Il est donc impératif d'assurer une surveillance humaine solide, en particulier dans les scénarios d'IA à haut risque. La surveillance humaine, souvent mise en œuvre au moyen de mécanismes "human-in-the-loop", a pour but d'assurer un contrôle et un équilibre dans les processus de prise de décision en matière d'IA, en veillant à ce que les décisions critiques impliquant des systèmes d'IA à haut risque restent sous le contrôle de l'homme.

Dans ce contexte, il est essentiel d'établir des dispositions juridiques solides qui garantissent véritablement la surveillance humaine, plutôt que de permettre des interprétations laxistes des exigences relatives à la présence humaine dans la boucle. Des études récentes ont mis en évidence les préoccupations concernant l'efficacité des dispositions relatives à la surveillance humaine, critiquant le fait que les êtres humains ne sont pas en mesure d'exercer les fonctions de surveillance souhaitées et que, par conséquent, ils pourraient indirectement légitimer l'utilisation de systèmes controversés.[[63]](#footnote-63) Nous recommandons donc que les États-Unis adoptent une approche rigoureuse centrée sur l'être humain. Cela pourrait impliquer la mise en œuvre de réglementations exigeant l'intervention humaine à des stades critiques du processus de prise de décision en matière d'IA, afin de garantir que le contrôle humain est activement et efficacement intégré au fonctionnement du système.

Compte tenu des limites actuelles de l'explicabilité, l'obligation d'adopter une approche centrée sur l'homme présente plusieurs avantages. Elle offre une protection contre les préjudices potentiels découlant de systèmes d'IA opaques, tout en favorisant la confiance et la responsabilité. En confiant aux humains le contrôle des décisions critiques, elle garantit une plus grande responsabilité et atténue les risques associés aux systèmes d'IA.

le manque de transparence. En outre, cette approche s'aligne sur les principes éthiques, en donnant la priorité aux valeurs et aux droits de l'homme dans le déploiement des technologies de l'IA. En fait, en imposant des obligations contraignantes aux HRS dans le cadre d'une approche centrée sur l'être humain, les États-Unis peuvent démontrer leur engagement en faveur d'un déploiement responsable de l'IA tout en répondant aux préoccupations pressantes concernant l'opacité et les risques liés à l'IA. Cette recommandation souligne l'importance de donner la priorité à l'action et à la surveillance humaines dans les processus de prise de décision en matière d'IA, en particulier dans les domaines où le risque de conséquences humaines est élevé.

1. **Établir des règles claires en matière de responsabilité afin de faciliter la réparation pour les personnes lésées par les systèmes d'IA**

Lors de la création de cadres réglementaires efficaces pour la gouvernance de l'IA, il est essentiel de trouver un équilibre entre les mesures préventives (ex ante) et les mécanismes de réparation des dommages (ex post). Historiquement, les États-Unis se sont fortement appuyés sur le droit ex post, en particulier dans la réglementation du numérique et de la technologie. Toutefois, pour minimiser les risques associés aux systèmes d'IA, nous préconisons une combinaison de dispositions ex ante et ex post. Parallèlement aux mesures visant à atténuer les risques avant qu'ils ne se produisent, il est urgent d'établir des règles de responsabilité solides afin d'offrir un recours aux personnes lésées par les systèmes d'IA.

Un élément clé de ces règles de responsabilité est l'introduction d'une "présomption de causalité". Cette disposition permettrait aux demandeurs d'indemnisation pour les dommages causés par les systèmes d'IA de bénéficier d'une charge de la preuve plus raisonnable et d'augmenter les chances de succès des actions en responsabilité. S'inspirant de la proposition de directive sur la responsabilité en matière d'IA, qui vise à adapter les règles de la responsabilité civile non contractuelle à l'IA, cette présomption de causalité transférerait une partie de la charge du demandeur au développeur ou à l'exploitant du système d'IA. En présumant l'existence d'un lien de causalité entre les actions du système d'IA et le préjudice qui en résulte, cette approche simplifie le processus juridique pour les victimes qui cherchent à obtenir réparation.

L'établissement de règles claires en matière de responsabilité permet non seulement aux individus de demander une indemnisation pour les dommages liés à l'IA, mais sert également de moyen de dissuasion contre le déploiement négligent ou imprudent des technologies de l'IA. En tenant les développeurs et les opérateurs responsables des conséquences de leurs systèmes d'IA, cette approche encourage l'innovation responsable à des stades plus précoces du développement des produits et favorise une plus grande confiance dans les technologies d'IA. En outre, elle s'inscrit dans le cadre d'efforts plus larges visant à faire respecter les droits fondamentaux et les principes éthiques dans le développement et le déploiement des systèmes d'IA.

En conclusion, l'établissement de règles claires en matière de responsabilité est essentiel pour garantir la responsabilité et faciliter la réparation pour les personnes lésées par les systèmes d'IA. En combinant des mesures *ex ante* avec des dispositions *ex post* solides, les décideurs politiques peuvent créer un cadre réglementaire qui encourage l'innovation tout en protégeant contre les risques et les préjudices potentiels associés aux technologies de l'IA.

1. **Création d'un groupe de travail de la Commission fédérale du commerce (FTC) chargé de superviser la mise en œuvre de l'intelligibilité de l'IA**

Afin de garantir la mise en œuvre effective des recommandations susmentionnées et de favoriser la transparence, la responsabilité et la confiance dans les systèmes d'IA, nous recommandons vivement la création d'un groupe de travail spécialisé au sein de la Commission fédérale du commerce des États-Unis (FTC). En tant que principale agence fédérale chargée de la protection des consommateurs et de l'application des règles de concurrence, la FTC est bien placée pour jouer un rôle de premier plan dans la supervision des pratiques de gouvernance de l'IA. Ce groupe de travail serait chargé de superviser la gouvernance et la réglementation de l'IA, en mettant l'accent sur la mise en œuvre des mesures proposées pour lutter contre l'opacité de l'IA et atténuer les risques qui y sont associés. Composé d'experts en éthique, en droit, en politique et en technologie de l'IA, ce groupe de travail contrôlerait le respect des cadres réglementaires, évaluerait l'efficacité des mesures de transparence et se pencherait sur les défis émergents en matière de gouvernance de l'IA.

En créant un groupe de travail spécialisé, la FTC peut assurer la surveillance et l'orientation nécessaires pour garantir que les systèmes d'IA sont développés et déployés d'une manière qui donne la priorité à l'explicabilité lorsque cela est nécessaire. En outre, le groupe de travail peut faciliter la collaboration entre les parties prenantes, y compris les agences gouvernementales, les acteurs de l'industrie, les universités et la société civile, afin de favoriser une approche multipartite de la gouvernance de l'IA et de la recherche sur son développement.

**Mesures d'accompagnement**

***Investissement dans la R&D sur l'explicabilité de l'IA***

Comme indiqué précédemment, le stade naissant de l'explicabilité de l'IA pose un problème en raison du manque de consensus et de clarté, qui empêche une mise en œuvre juridique efficace. Pour y remédier, il faut investir massivement dans la recherche et le monde universitaire. Une première étape importante a déjà été franchie avec la création du Centre national de recherche sur l'IA (NAIRR).[[64]](#footnote-64) Le NAIRR, doté d'un budget de 2,6 milliards de dollars pour les six prochaines années, a pour objectif central de promouvoir une "IA digne de confiance". Toutefois, la majeure partie du financement, qui s'élève à 2,25 milliards de dollars, sera allouée aux fournisseurs de ressources du NAIRR, qui comprennent principalement des entreprises privées et des agences fédérales.[[65]](#footnote-65) Il n'est donc pas certain que ces ressources soutiendront efficacement la recherche sur l'explicabilité de l'IA. Il est surprenant de constater que les termes "explicable" et "expliquable" sont absents du rapport NAIRR 2023. Par conséquent, nous recommandons que le financement du NAIRR donne la priorité à la recherche sur l'IA explicable ou que le financement de l'innovation en matière d'IA soit conditionné au développement de systèmes d'IA explicables.

En outre, les États-Unis devraient encourager la collaboration internationale en matière de recherche. Cela pourrait se faire en prolongeant l'arrangement administratif sur "l'IA pour le bien public" signé par l'UE et les États-Unis en janvier 2023.[[66]](#footnote-66) Des cadres similaires devraient être proposés, axés sur le financement et la promotion de partenariats de recherche internationaux pour faire progresser l'explicabilité de l'IA. Par exemple, un accord similaire pourrait être proposé avec la Chine, promouvant ainsi le dialogue sur l'IA entre les États-Unis et la Chine qui a déjà commencé à un niveau administratif.[[67]](#footnote-67) En investissant dans des initiatives de recherche collaborative, nous pouvons accélérer le développement de techniques et de normes, réduisant ainsi l'ambiguïté de la gouvernance de l'IA. En outre, il convient de mettre l'accent sur la promotion des collaborations de recherche interdisciplinaires entre les informaticiens, les éthiciens, les psychologues et d'autres domaines pertinents. Cette approche permet d'explorer les diverses dimensions de l'explicabilité de l'IA et de développer des solutions globales.

***Création d'un groupe de travail international sur les normes d'explicabilité de l'IA***

Les États-Unis devraient être le fer de lance de la création d'un groupe de travail international sur les normes d'explicabilité de l'IA, en s'appuyant sur les efforts de collaboration existants, tels que le Conseil du commerce et de la technologie UE-États-Unis.[[68]](#footnote-68) Ce groupe de travail se concentrerait sur l'élaboration et l'harmonisation de normes internationales en matière d'explicabilité de l'IA. En tirant parti des partenariats et de l'expertise de diverses nations, il vise à établir des définitions, des cadres, des normes et des meilleures pratiques communs. Au sein de ce groupe de travail, les propositions d'éthique de l'IA de différents pays (documents du NIST aux États-Unis, documents du ministère chinois de la science, dispositions de la loi européenne sur l'IA) peuvent être comparées, ce qui pourrait déboucher sur un consensus international et une coopération en matière de réglementation de l'IA. Ce groupe de travail pourrait être composé de membres d'organisations internationales qui ont contribué à la recherche sur l'explicabilité de l'IA, telles que l'OCDE et l'ISO. En participant activement à cet effort de collaboration, les États-Unis pourraient faire preuve de leadership en faisant progresser les efforts mondiaux en faveur d'un déploiement de l'IA digne de confiance et responsable.

# **Conclusion**

En naviguant dans les complexités de l'explicabilité de l'IA, les décideurs politiques sont confrontés à un délicat exercice d'équilibre. L'impératif de transparence, de responsabilité et de confiance de la société dans les technologies de l'IA exige des cadres réglementaires solides qui répondent efficacement aux défis posés par les processus décisionnels opaques. Toutefois, pour atteindre cet objectif, il convient d'adopter une approche nuancée qui tienne compte des complexités techniques des systèmes d'IA tout en protégeant contre les risques et les préjudices potentiels.

Cette note politique propose un examen complet de l'explicabilité de l'IA, en s'appuyant sur les connaissances académiques et les approches réglementaires internationales pour fournir des recommandations applicables aux décideurs politiques américains. En synthétisant les trois principales approches de la mise en œuvre de l'explicabilité et en distillant des principes clés, ce travail plaide en faveur de la nécessité d'une approche réglementaire qui donne la priorité aux impératifs éthiques tout en tenant compte des contraintes pratiques.

Les principales recommandations que nous proposons, qui résultent de l'analyse des options politiques, sont les suivantes :

* L'**adoption d'une approche fondée sur les risques**
* Imposer des **obligations contraignantes pour les systèmes à haut risque** en mettant l'**accent sur l'être humain**
* Établir des **règles claires en matière de responsabilité** afin de faciliter la réparation pour les personnes lésées par l'IA
* La formation d'un groupe de travail spécialisé au sein de la FTC pour superviser la gouvernance de l'IA.

En suivant les recommandations présentées dans ce document, les décideurs politiques américains peuvent naviguer efficacement dans les complexités de la gouvernance de l'IA, en veillant à ce que les technologies de l'IA soient déployées d'une manière qui respecte les valeurs sociétales, favorise la confiance et promeut l'innovation responsable au profit de tous.

Dans le même temps, nous reconnaissons certaines limites qui entravent la mise en œuvre de l'explicabilité dans l'IA. La complexité technique des systèmes d'IA, en particulier des réseaux neuronaux profonds et d'autres systèmes d'IA avancés, constitue un obstacle important à la mise en œuvre de mesures concrètes en faveur de la transparence. En outre, la nature évolutive de la technologie de l'IA constitue une cible mouvante pour les décideurs politiques, ce qui nécessite une approche souple en matière de réglementation et d'application.

Pour l'avenir, il est essentiel de poursuivre les recherches pour répondre au besoin pressant de solutions techniques permettant de mettre en œuvre efficacement l'explicabilité dans la gouvernance de l'IA. Par conséquent, la recherche subventionnée par les pouvoirs publics doit être consacrée au développement de techniques et de méthodologies innovantes qui améliorent l'interprétabilité et la transparence des systèmes d'IA.

#

#

#

**Bibliographie**

Angwin, Julia, Jeff Larson, Surya Mattu et Lauren Kirchner. "Machine Bias". ProPublica, 23 mai 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bibal, A., Lognoul, M., de Streel, A. et al. "Legal requirements on explainability in machine learning" *Artificial Intelligence Law* 29, 149-169 (2021). <https://doi.org/10.1007/s10506-020-09270-4>

CSET "Traduction : Normes éthiques pour la nouvelle génération d'intelligence artificielle diffusée". Université de Georgetown. (21 octobre 2021). Consulté le 2 avril 2024 sur <https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/>

Ebers, Martin. "Réglementer l'IA explicable dans l'Union européenne. An Overview of the Current Legal Framework(S)." papers.ssrn.com. Rochester, NY, 9 août 2021. https://ssrn.com/abstract=3901732.

CEPD. "TechDispatch #2/2023 - Explainable Artificial Intelligence". (16 novembre 2023). Consulté le 3 avril 2024 sur <https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en>

ÉQUIPE ETSI COMS. "Le groupe Securing AI de l'ETSI devient un comité technique pour aider l'ETSI à répondre à la loi européenne sur l'IA. ETSI, 17 octobre 2023. https://www.etsi.org/newsroom/news/2288-etsi-s-securing-ai-group-becomes-a-technical-committee-to-help-etsi-to-answer-the-eu-ai-act.

Commission européenne. Communication intitulée "Favoriser une approche européenne de l'intelligence artificielle - Façonner l'avenir numérique de l'Europe", digital-strategy.ec.europa.eu, 21 avril 2021. https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence.

"Plate-forme ENorm", ec.europa.eu, 22 mai 2023. https://ec.europa.eu/growth/tools-databases/enorm/mandate/593\_en.

Groupe d'experts de haut niveau sur l'IA. "Lignes directrices en matière d'éthique pour une IA digne de confiance". (2018). Consulté le 31 mars 2024 sur https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

"Liability Rules for Artificial Intelligence", commission.europa.eu, 2022. https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence\_en.

"TTC Joint Roadmap for Trustworthy AI and Risk Management | Shaping Europe's Digital Future." digital-strategy.ec.europa.eu, December 2, 2022. https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management.

Parlement européen. "Loi sur l'intelligence artificielle, 2019. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138\_EN.pdf.

Service de recherche du Parlement européen. "Approche des États-Unis en matière d'intelligence artificielle". Epthinktank, 18 janvier 2024. https://epthinktank.eu/2024/01/18/united-states-approach-to-artificial-intelligence/.

Décret exécutif. No. 14110, 3 C.F.R. 88 § (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al : "AI4People - un cadre éthique pour une bonne société de l'IA : opportunités, risques, principes et recommandations". *Minds & Machines* 28, 689-707 (2018). [https://doi.](https://doi.org/10.1007/s11023-018-9482-5)org/10.1007/s11023-018-9482-5

G20. "ANNEXE Principes du G20 en matière d'IA, 2019. https://www.mofa.go.jp/policy/economy/g20\_summit/osaka19/pdf/documents/en/annex\_08.pdf.

Goddard, Kate, Abdul Roudsari, et Jeremy C Wyatt. "Automation Bias : A Systematic Review of Frequency, Effect Mediators, and Mitigators" (Biais d'automatisation : examen systématique de la fréquence, des médiateurs d'effets et des atténuateurs). *Journal of the American Medical Informatics Association* 19, no. 1 (janvier 2012) : 121–27. htt[ps://doi.or](https://doi.org/10.1136/amiajnl-2011-000089)g/10.1136/amiajnl-2011-000089.

Gouvernement de la Chine "Loi sur la protection des informations personnelles". Adoptée lors de la 30e réunion du Comité permanent de la 13e Assemblée populaire nationale. (20 août 2021)

Green, Ben. "The flaws of policies requiring human oversight of government algorithms" (Les failles des politiques exigeant un contrôle humain des algorithmes gouvernementaux). *Computer Law & Security Review* 45 (2022). Disponible à l'adresse [:](https://arxiv.org/ftp/arxiv/papers/2109/2109.05067.pdf) https://arxiv.org/ftp/arxiv/papers/2109/2109.05067.pdf

Hagendorff, T. "L'éthique de l'IA : An Evaluation of Guidelines". *Minds & Machines* 30, 99-120 (2020). [https://doi.](https://doi.org/10.1007/s11023-020-09517-8)org/10.1007/s11023-020-09517-8

Herzog, C. "On the risk of confusing interpretability with explicability". *AI Ethics* 2, 219-225 (2022). [https://doi.](https://doi.org/10.1007/s43681-021-00121-9)org/10.1007/s43681-021-00121-9

" How Dutch activists got an invasive fraud detection algorithm banned ", AlgorithmWatch, consulté le 6 avril 2024, https://algorithmwatch.org/en/syri-netherlands-algorithm/.

ISO. "ISO/IEC JTC 1/SC 42 - Intelligence artificielle". ISO, 2017. https://www.iso.org/committee/6794475.html.

Kaminski, Margot, Andrea Bertolini, Kiel Brennan-Marquez, Giovanni Comandé, Matthew Cushing, Natalie Helberger, Max Van Drunen, et al. "THE RIGHT to EXPLANATION, EXPLAINED." *BERKELEY TECHNOLOGY LAW JOURNAL* 34 (2019) : 189. https[://doi.org/](https://doi.org/10.15779/Z38TD9N83H)10.15779/Z38TD9N83H.

Kalmykov, V. L. "XXAI : Explicitly Explainable AI provides transparency in automatic decision-making by overcoming limitations of symbolic AI[". https://doi.org/10.48550/arXiv.2401.03093](https://doi.org/10.48550/arXiv.2401.03093)

Lantham & Watkins "China's New AI Regulations". Latham & Watkins Privacy & Cyber Practice, Client Alert Commentary, No 3110, (16 août 2023).

Lewis, James Andrew, Benson, Emily et Michael Frank, "The Biden Administration's Executive Order on Artificial Intelligence", Commentaire, 31 octobre 2023, Centre for Strategic and International Studies. Consulté le 6 avril à l'adresse suivante : https://www.csis.org/analysis/biden-administrations-executive-order-artificial-intelligence

Mohammadi, B., Malik, N., Derdenger, T., Srinivasan, K. "Regulation eXplainable Artificial Intelligence (XAI) May Harm Consumers, arXiv (2022). <https://doi.org/10.48550/arXiv.2209.03499>

Groupe de travail NAIRR. "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem an Implementation Plan for a National Artificial Intelligence Research Resource," 2023. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.

Nannini, Luca, Agathe Balayn, et Adam Leon Smith. "Explainability in AI Policies : A Criticial Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK". *ACM*, 12 juin 2023.

NIST. "Cadre de gestion des risques liés à l'IA. NIST, 12 juillet 2021. https://www.nist.gov/itl/ai-risk-management-framework.

OCDE. "Quarante-deux pays adoptent les nouveaux principes de l'OCDE sur l'intelligence artificielle - OCDE", www.oecd.org, 2019. https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm.

"Portail de l'Observatoire des politiques d'IA de l'OCDE" oecd.ai, n.d. https://oecd.ai/en/dashboards/ai-principles/P7.

"Les principes de l'OCDE en matière d'intelligence artificielle (IA)." oecd.ai, 2019. https://oecd.ai/en/ai-principles.

Panigutti, Cecilia, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, et al. "The Role of Explainable AI in the Context of the AI Act". *2023 ACM Conference on Fairness, Accountability, and Transparency*, 12 juin 2023. https://doi.org/10.1145/3593013.3594069.

Phillips, P Jonathon, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, et Mark A Przybocki. "Four Principles of Explainable Artificial Intelligence (Quatre principes d'intelligence artificielle explicable). *National Institute of Standards and Technology*, 29 septembre 2021. https://doi.org/10.6028/nist.ir.8312.

Reddy, G. Pradeep et Y.V. Pavan Kumar. "Explainable AI (XAI) : Explained | IEEE Conference Publication | IEEE Xplore." ieeexplore.ieee.org, 27 avril 2023. https://ieeexplore.ieee.org/abstract/document/10134984.

Règlement (UE) 2016/679, Règlement général sur la protection des données, Eur., 2016. Consulté le 31 mars 2024 sur https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Robbins, Scott. "A Misdirected Principle with a Catch : Explicability for AI". *Minds and Machines* 29, no. 4 (15 octobre 2019) : 495–514. htt[ps://doi.or](https://doi.org/10.1007/s11023-019-09509-3)g/10.1007/s11023-019-09509-3.

Rolf, S. "China's Regulations on Algorithms. Contexte, impact et comparaisons avec l'UE". *FES Briefing,* Friedrich Ebert Stiftung (janvier 2023).

Schuett, Jonas. "Risk Management in the Artificial Intelligence Act" (Gestion des risques dans la loi sur l'intelligence artificielle). *European Journal of Risk Regulation*, 8 février 2023, 1-19. https://doi.org/10.1017/err.2023.1.

Sheehan, M. "China's AI Regulations and How They Get Made". Fondation Carnegie pour la paix internationale (10 juillet 2023).

Sheehan, M. "Tracing the Roots of China's AI Regulations". Carnegie Endowment of International Peace (27 février 2024).

Sovrano, Francesco, Salvatore Sapienza, Monica Palmirani et Fabio Vitali. "Metrics, Explainability and the European AI Act Proposal". *J* 5, no. 1 (18 février 2022) : 126-38. https://doi.org/10.3390/j5010010.

Siegmann, Charlotte, et Markus Anderljung. "L'effet Bruxelles et l'intelligence artificielle : How EU regulation will impact the global AI market." arXiv *preprint arXiv:2208.12645* (2022).

Tabassi, Elham. "Cadre de gestion des risques de l'intelligence artificielle (AI RMF 1.0). *National Institute of Standards and Technology*, 26 janvier 2023. https://doi.org/10.6028/nist.ai.100-1.

The Royal Society "L'IA explicable : les bases". (novembre 2019) ISBN : 978-1-78252-433-5.

La Maison Blanche. "FACT SHEET : Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety (Fiche d'information : l'administration Biden-Harris annonce de nouvelles actions pour promouvoir une innovation responsable en matière d'IA qui protège les droits et la sécurité des Américains). Maison Blanche, 4 mai 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

Congrès américain. Algorithmic Accountability Act of 2022, S.3572 § (2022). [https://www](https://www.congress.gov/bill/117th-%20congress/senate-%20bill/3572/text).congress.gov/bill/117th-%20congress/senate-%20bill/3572/text...

Webster, Graham , et Ryan Hass. "A Roadmap for a US-China AI Dialogue" (Feuille de route pour un dialogue sur l'IA entre les États-Unis et la Chine). Brookings, 10 janvier 2024. https://www.brookings.edu/articles/a-roadmap-for-a-us-china-ai-dialogue/.

Wachter, Sandra, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, Volume 7, Issue 2, May 2017, Pages 76-99, [https://doi.org/10.1093/idpl/ipx005.](https://doi.org/10.1093/idpl/ipx005)

**A propos des auteurs :**

**Dominique Boullier** a une formation de sociologue et de linguiste. Il est professeur des Universités en sociologie à l'Institut d'études politiques de Paris (Sciences Po) depuis 2009 (émérite depuis 2022). Chercheur au CEE (Centre d'études européennes et de politique comparée). Ancien directeur du Social Media Lab de l'École polytechnique fédérale de Lausanne (EPFL) (2015-2019). Ancien coordinateur scientifique du Medialab à Sciences Po avec Bruno Latour (2009-2013). Ancien directeur du Lutin User Lab (Cité des Sciences, Paris (2004-2008) et de Costech (Université de Technologie de Compiègne (1998-2005). Ancien entrepreneur dans les années 90.

Il est l'auteur d'articles et d'ouvrages sur les questions urbaines (L'urbanité numérique, L'Harmattan, 1999 ; La ville-événement, PUF, 2010), sur les technologies numériques (Sociologie du numérique, Armand Colin, 2019, 2ème édition). Ses recherches actuelles portent sur les conditions techniques et institutionnelles de la survie dans le monde numérique : « Comment sortir de l'emprise des réseaux sociaux ? “ (Le Passeur éditeur, 2020), ” Puissance des plateformes. Territoires et souverainetés » (Sciences Po Chaire digital, 2021), “Habiter le numérique. L'habitèle", (à paraître) et sur les mutations conceptuelles et méthodologiques nécessaires aux sciences sociales : « Propagations. Un nouveau paradigme pour les sciences sociales » (Armand Colin, 2023).

**A propos de la chaire Digital, gouvernance et souveraineté :**

La [**Chaire Digital, Gouvernance et Souveraineté de Sciences Po**](https://www.sciencespo.fr/public/chaire-numerique/) a pour mission de favoriser un forum unique réunissant des entreprises techniques, des universitaires, des décideurs politiques, des acteurs de la société civile, des incubateurs de politiques publiques ainsi que des experts de la régulation numérique.

Hébergée par [**l'École d'affaires publiques**](https://www.sciencespo.fr/ecole-affaires-publiques/fr/), la Chaire adopte une approche pluridisciplinaire et holistique pour rechercher et analyser les transformations économiques, juridiques, sociales et institutionnelles induites par l'innovation numérique. La Chaire Digital, Gouvernance et Souveraineté est présidée par **Florence G'sell**, professeur de droit à l'Université de Lorraine, maître de conférences à l'École d'affaires publiques de Sciences Po. Durant l'année universitaire 2023-2024, Florence G'sell est professeur invité au Cyber Policy Center de l'Université de Stanford.

*Les activités de la Chaire sont soutenues par :*

 

1. Julia Angwin et al, "Machine Bias", ProPublica, 23 mai 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [↑](#footnote-ref-1)
2. L. Floridi et al, "AI4People - un cadre éthique pour une bonne société de l'IA : opportunités, risques, principes et recommandations", *Minds & Machines,* vol. 28 (2018). [↑](#footnote-ref-2)
3. V. L. Kalmykov, "XXAI : Explicitly Explainable AI provides transparency in automatic decision-making by overcoming the limitations of symbolic AI," *arXiv*, (2024). [↑](#footnote-ref-3)
4. V. L. Kalmykov, 2024. [↑](#footnote-ref-4)
5. Victor Demiaux, " Comment l'humain peut-il garder la main ? ", CNIL, décembre 2017, https://www.cnil.fr/sites/cnil/files/atoms/files/cnil\_rapport\_ai\_gb\_web.pdf. [↑](#footnote-ref-5)
6. L. Floridi et al, 2018. [↑](#footnote-ref-6)
7. Ibid. [↑](#footnote-ref-7)
8. L. Floridi et al, 2018. [↑](#footnote-ref-8)
9. C. Herzog, "On the risk of confusing interpretability with explicability", *AI Ethics,* vol. 2 (2022). [↑](#footnote-ref-9)
10. Ibid. [↑](#footnote-ref-10)
11. L. Floridi et al, "AI4People - un cadre éthique pour une bonne société de l'IA : opportunités, risques, principes et recommandations", *Minds & Machines,* vol. 28 (2018). [↑](#footnote-ref-11)
12. Charlotte Siegmann et Markus Anderljung, "The Brussels effect and artificial intelligence : How EU regulation will impact the global AI market," arXiv *preprint arXiv:2208.12645* (2022). [↑](#footnote-ref-12)
13. OCDE, "Les principes de l'OCDE en matière d'intelligence artificielle (IA)", oecd.ai, 2024, https://oecd.ai/en/ai-principles. [↑](#footnote-ref-13)
14. Ibid. [↑](#footnote-ref-14)
15. OCDE, "Portail de l'Observatoire des politiques d'IA de l'OCDE", oecd.ai, s.d., https://oecd.ai/en/dashboards/ai-principles/P7. [↑](#footnote-ref-15)
16. Ibid. [↑](#footnote-ref-16)
17. G20, "ANNEXE G20 AI Principles", 2019, https://www.mofa.go.jp/policy/economy/g20\_summit/osaka19/pdf/documents/en/annex\_08.pdf. [↑](#footnote-ref-17)
18. OCDE, "Quarante-deux pays adoptent les nouveaux principes de l'OCDE sur l'intelligence artificielle - OCDE", www.oecd.org, 2019, https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm. [↑](#footnote-ref-18)
19. Commission européenne, "ENorm Platform", ec.europa.eu, 22 mai 2023, https://ec.europa.eu/growth/tools-databases/enorm/mandate/593\_en. [↑](#footnote-ref-19)
20. ETSI COMS TEAM, "ETSI's Securing AI Group Becomes a Technical Committee to Help ETSI to Answer the EU AI Act", ETSI, 17 octobre 2023, https://www.etsi.org/newsroom/news/2288-etsi-s-securing-ai-group-becomes-a-technical-committee-to-help-etsi-to-answer-the-eu-ai-act. [↑](#footnote-ref-20)
21. NIST, "AI Risk Management Framework", NIST, 12 juillet 2021, https://www.nist.gov/itl/ai-risk-management-framework. [↑](#footnote-ref-21)
22. ISO, "ISO/IEC JTC 1/SC 42 - Intelligence artificielle", ISO, 2017, https://www.iso.org/committee/6794475.html. [↑](#footnote-ref-22)
23. G. Pradeep Reddy et Y.V. Pavan Kumar, "Explainable AI (XAI) : Explained | IEEE Conference Publication | IEEE Xplore," ieeexplore.ieee.org, 27 avril 2023, https://ieeexplore.ieee.org/abstract/document/10134984. [↑](#footnote-ref-23)
24. Règlement (UE) 2016/679, Règlement général sur la protection des données, Eur., 2016. [↑](#footnote-ref-24)
25. Margot Kaminski et al, "THE RIGHT to EXPLANATION, EXPLAINED", BERKELEY TECHNOLOGY LAW JOURNAL 34 (2019) : 189, https[://doi.org/](https://doi.org/10.15779/Z38TD9N83H)10.15779/Z38TD9N83H. [↑](#footnote-ref-25)
26. Sandra Wachter, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, Volume 7, Issue 2, May 2017. <https://doi.org/10.1093/idpl/ipx005> [↑](#footnote-ref-26)
27. Martin Ebers, Réglementer l'IA explicable dans l'Union européenne. An Overview of the Current Legal Framework(s) (9 août 2021). Liane Colonna/Stanley Greenstein (eds.), Nordic Yearbook of Law and Informatics 2020 : Law in the Era of Artificial Intelligence, disponible sur SSRN [:](https://dx.doi.org/10.2139/ssrn.3901732) http://dx.doi.org/10.2139/ssrn.3901732 [↑](#footnote-ref-27)
28. Commission européenne, "Communication sur la promotion d'une approche européenne de l'intelligence artificielle - Façonner l'avenir numérique de l'Europe", digital-strategy.ec.europa.eu, 21 avril 2021, https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence. [↑](#footnote-ref-28)
29. Cecilia Panigutti et al, "The Role of Explainable AI in the Context of the AI Act", 2023 ACM Conference on Fairness, Accountability, and Transparency, 12 juin 2023, https://doi.org/10.1145/3593013.3594069. [↑](#footnote-ref-29)
30. Parlement européen, "Artificial Intelligence Act", 2019, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138\_EN.pdf. [↑](#footnote-ref-30)
31. Francesco Sovrano et al, "Metrics, Explainability and the European AI Act Proposal", J 5, no 1 (18 février 2022) : 126-38, https://doi.org/10.3390/j5010010. [↑](#footnote-ref-31)
32. Parlement européen, "Artificial Intelligence Act", 2024 [↑](#footnote-ref-32)
33. Jonas Schuett, "Risk Management in the Artificial Intelligence Act", European Journal of Risk Regulation, 8 février 2023, 1-19, https://doi.org/10.1017/err.2023.1. [↑](#footnote-ref-33)
34. Pieter Haeck, "Robo-Cop : EU Wants Firms to Be Held Liable for Harmone Done by AI", Politico, 28 septembre 2022, https://www.politico.eu/article/artificial-intelligence-european-commission-ai-liability-directive/. [↑](#footnote-ref-34)
35. Commission européenne, "Liability Rules for Artificial Intelligence", commission.europa.eu, 2022, https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence\_en. [↑](#footnote-ref-35)
36. Luca Nannini, Agathe Balayn, et Adam Leon Smith, "Explainability in AI Policies : A Criticial Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK", ACM, 12 juin 2023. [↑](#footnote-ref-36)
37. S. Rolf, "China's Regulations on Algorithms. Context, Impact, and comparisons with the EU", *FES Briefing*, Friedrich Ebert Stiftung, janvier 2023. [↑](#footnote-ref-37)
38. Latham & Watkins, 2023. [↑](#footnote-ref-38)
39. CSET, "Traduction : Ethical Norms for New Generation Artificial Intelligence Released". Georgetown University, 21 octobre 2021, Consulté le 2 avril 2024 sur https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/. [↑](#footnote-ref-39)
40. M. Sheehan, "China's AI Regulations and How They Get Made", Carnegie Endowment for International Peace, 10 juillet 2023. [↑](#footnote-ref-40)
41. Gouvernement de la Chine, "Loi sur la protection des informations personnelles", adoptée lors de la 30e réunion du Comité permanent du 13e Congrès national du peuple, le 20 août 2021. [↑](#footnote-ref-41)
42. M. Sheehan, "China's AI Regulations and How They Get Made", Carnegie Endowment for International Peace, 10 juillet 2023. [↑](#footnote-ref-42)
43. M. Sheehan, "Tracing the Roots of China's AI Regulations", Carnegie Endowment of International Peace, 27 février 2024. [↑](#footnote-ref-43)
44. Latham & Watkins, "China's New AI Regulations", Latham & Watkins Privacy & Cyber Practice, Client Alert Commentary, No 3110, August 16 2023. [↑](#footnote-ref-44)
45. P Jonathon Phillips et al, "Four Principles of Explainable Artificial Intelligence", National Institute of Standards and Technology, 29 septembre 2021, https://doi.org/10.6028/nist.ir.8312. [↑](#footnote-ref-45)
46. Ibid. [↑](#footnote-ref-46)
47. Elham Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", National Institute of Standards and Technology, 26 janvier 2023, https://doi.org/10.6028/nist.ai.100-1. [↑](#footnote-ref-47)
48. Congrès américain, "Algorithmic Accountability Act of 2022", S.3572 § (2022), https://www.congress.gov/bill/117th-%20congress/senate-%20bill/3572/text. [↑](#footnote-ref-48)
49. Executive Order No. 14110, 3 C.F.R. 88 § (2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ . [↑](#footnote-ref-49)
50. Ibid. [↑](#footnote-ref-50)
51. Ibid. [↑](#footnote-ref-51)
52. Ibid. [↑](#footnote-ref-52)
53. B. Mohammadi et al, "Regulation eXplainable Artificial Intelligence (XAI) May Harm Consumers, arXiv (2022). [↑](#footnote-ref-53)
54. A. Bibal et al, "Legal requirements on explainability in machine learning", *Artificial Intelligence Law,* 29, 149-169 (2021) https://doi.org/10.1007/s10506-020-09270-4. [↑](#footnote-ref-54)
55. B. Mohammadi et al, "Regulation eXplainable Artificial Intelligence (XAI) May Harm Consumers, arXiv (2022). [↑](#footnote-ref-55)
56. CEPD, "TechDispatch #2/2023 - Explainable Artificial Intelligence". 16 novembre 2023, Consulté le 3 avril 2024 sur https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence\_en. [↑](#footnote-ref-56)
57. Ibid. [↑](#footnote-ref-57)
58. Ibid. [↑](#footnote-ref-58)
59. The Royal Society, "Explainable AI : the basics", novembre 2019, ISBN : 978-1-78252-433-5. [↑](#footnote-ref-59)
60. P Jonathon Phillips et al, 2021. [↑](#footnote-ref-60)
61. Panigutti et al, 2023. [↑](#footnote-ref-61)
62. James Andrew Lewis, Emily Benson et Michael Frank, "The Biden Administration's Executive Order on Artificial Intelligence", Commentaire, 31 octobre 2023, Centre for Strategic and International Studies. Consulté le 6 avril à l'adresse suivante : https://www.csis.org/analysis/biden-administrations-executive-order-artificial-intelligence [↑](#footnote-ref-62)
63. Ben Green, "The flaws of policies requiring human oversight of government algorithms". *Computer Law & Security Review* 45 (2022). [↑](#footnote-ref-63)
64. Maison Blanche, "FACT SHEET : Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety", Maison Blanche, 4 mai 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/. [↑](#footnote-ref-64)
65. Ibid. [↑](#footnote-ref-65)
66. Service de recherche du Parlement européen, "United States Approach to Artificial Intelligence", Epthinktank, 18 janvier 2024, https://epthinktank.eu/2024/01/18/united-states-approach-to-artificial-intelligence/. [↑](#footnote-ref-66)
67. Graham Webster et Ryan Hass, "A Roadmap for a US-China AI Dialogue", Brookings, 10 janvier 2024, https://www.brookings.edu/articles/a-roadmap-for-a-us-china-ai-dialogue/. [↑](#footnote-ref-67)
68. Commission européenne, "TTC Joint Roadmap for Trustworthy AI and Risk Management | Shaping Europe's Digital Future", digital-strategy.ec.europa.eu, 2 décembre 2022, https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management. [↑](#footnote-ref-68)