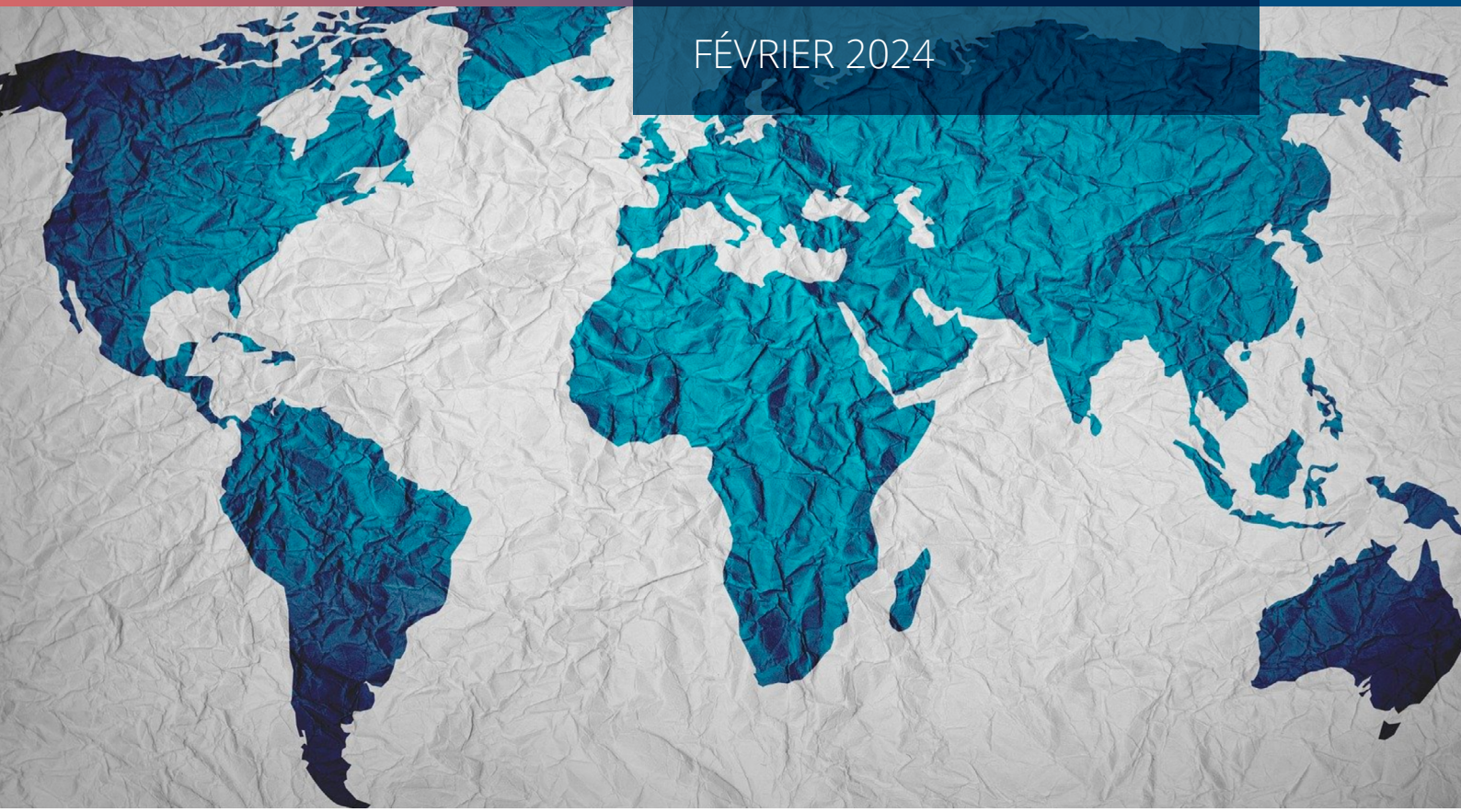
L'IA comme bien public : garantir un contrôle démocratique de l'IA dans l'espace informationnel

CADRE DE RÉGULATION

FÉVRIER 2024



Forum Information & Démocratie

SOMMAIRE

UNE FEUILLE DE ROUTE POUR L'IA DANS L'INTÉRÊT PUBLIC

PAR LAURA SCHERTEL MENDES & JONATHAN STRAY

Coprésidents du groupe de travail

Il y a six mois, nous avons été sollicités pour co-présider le groupe de travail sur l'intelligence artificielle et ses implications pour l'espace de l'information et de la communication, mis en place par le Forum sur l'information et la démocratie. Alors que nous nous engagions dans ce processus ambitieux visant à apporter une contribution substantielle et démocratique au débat mondial sur l'IA et sa réglementation, de nombreuses questions restaient sans réponse.

En tant qu'informaticien et journaliste, moi, Jonathan Stray, j'ai étudié l'influence des systèmes d'IA sur les résultats notamment sur l'information des citoyens, le bien-être des utilisateurs, et particulièrement la polarisation et les conflits. Avec mes collègues du Center for Human-compatible AI de l'UC Berkeley, nous étudions les problèmes qui se posent et les solutions existantes. Une énorme fraction des informations que toutes les personnes voient est maintenant sélectionnée par des algorithmes de plus en plus sophistiqués. Quels biais ces systèmes omniprésents ont-ils, et comment ce qu'ils nous montrent va-t-il aggraver ou désamorcer nos désaccords ? De plus, comment les acteurs malveillants peuvent-ils manipuler ces systèmes pour nous tromper ou nous diviser, par le biais des médias synthétiques ou de la persuasion en réseau ? Nous devons faire mieux que de laisser les choses se produire, en combinant des principes de conception prudents et une réglementation qui garantisse la prise en compte de l'intérêt public. Nous avons également besoin de mécanismes politiques permettant aux chercheurs externes d'étudier le fonctionnement interne des systèmes commerciaux, afin que nous puissions tous comprendre ce que ces machines extrêmement influentes font réellement aux individus et aux sociétés.

En tant que professeure de droit, moi, Laura Schertel Mendes, j'ai été fortement impliquée dans la traduction des principes éthiques de l'IA en réglementation. En tant que rapporteure de la Commission brésilienne des juristes, j'ai conseillé le Sénat brésilien sur l'approche à adopter pour réglementer l'IA, élaborant ainsi le premier projet de loi sur l'IA au Brésil. Nous visons à mettre en place une gouvernance globale de l'IA pour garantir que les technologies d'IA sont développées et utilisées de manière responsable et éthique, tout en minimisant les risques et les dommages potentiels. Nous voulons créer, par le biais de la réglementation, des mécanismes de responsabilité externes aux entreprises et accessibles à la société. Nous reconnaissons que cela peut prendre différentes formes dans différents pays, pour répondre aux divers contextes culturels, comme nous l'avons discuté dans notre Groupe de travail international.

Ce rapport est le résultat de six mois de discussions intensives et inspirantes, testant des idées avec des experts de diverses disciplines et horizons du monde entier. Nos réunions de groupe de travail comprenaient des séances dédiées sur un large éventail de sujets, y compris les régimes de responsabilité, les mesures techniques d'atténuation, les normes d'authenticité et les options politiques. Ces discussions nous ont permis d'élaborer une feuille de route visant à garantir que l'IA soit développée, déployée et utilisée dans l'intérêt public. En considérant l'IA comme un bien public, nous plaidons en faveur d'un changement de priorités. Les systèmes d'IA doivent être sûrs, équitables et fiables si l'on veut qu'ils constituent une innovation bénéfique pour tous les citoyens et qu'ils favorisent un développement durable.

La gouvernance démocratique de l'espace de l'information et de la communication est une condition préalable à la démocratie elle-même. L'utilisation responsable de l'IA peut promouvoir cette gouvernance démocratique ou, au contraire, l'atténuer. La série de mesures proposées dans ce cadre de régulation permet le contrôle démocratique de l'IA dans l'espace informationnel.

Nos recommandations s'adressent à des acteurs tels que les développeurs d'IA, les déployeurs et les gouvernements, les appelant à mettre en œuvre des mesures concrètes et réalisables pour construire des systèmes d'IA plus éthiques, inclusifs et responsables. Les personnes et les organisations qui alimentent la révolution de l'IA peuvent être à l'avant-garde des développements technologiques au service des intérêts des citoyens.

Les recommandations sont formulées de manière flexible afin qu'elles puissent être mises en œuvre par différents acteurs dans des contextes et des pays très différents. En même temps, en réponse à la demande de lignes directrices qui vont au-delà des principes généraux pour guider la pratique, elles sont concrètes et proposent des mesures spécifiques.

Nous plaçons les citoyens au cœur de nos préoccupations, en appelant à une consultation proactive, à des systèmes de responsabilité et à des mécanismes de recours. Nous combinons une approche de réglementation de l'IA fondée sur les risques avec une approche basée sur les droits qui garantit le droit à l'information, le droit de ne pas être discriminé, le droit de recevoir une explication et le droit de contester un résultat généré par une machine.

Ceux d'entre nous qui sont des chercheurs et des acteurs de la société civile jouent un rôle particulièrement important dans l'étude de ces systèmes, les examinant pour vérifier leur conformité mais aussi pour éclairer l'élaboration des politiques publiques et le développement de systèmes plus fiables. Pourtant, les développeurs d'IA sont réticents à fournir l'accès à leurs systèmes et aux données d'entraînement. Ce n'est que par une réglementation appropriée que nous pouvons permettre aux parties prenantes externes d'apporter leur contribution à un développement technologique au service de la démocratie.

Nous tenons à remercier nos collègues du Groupe de travail pour leurs précieuses idées et contributions, tous les nombreux experts qui ont donné de leur temps et de leurs connaissances, ainsi que les efforts extraordinaires de l'équipe qui a dirigé et rédigé ce cadre de régulation : Katharina Zügel, Viviana Padelli, Lia Chkhetiani et Kaye Celine Palisoc.

Si le développement et l'utilisation de l'IA se poursuivent comme ils le font actuellement, ils posent des défis majeurs à l'environnement informationnel qui alimente les processus démocratiques. En tant qu'Américain et Brésilienne, nous regardons avec inquiétude les prochaines élections dans nos pays. Nous sommes au bord d'un changement majeur dans le paysage de la gouvernance de l'IA, passant des idées à la réglementation. Il est temps que les États agissent, et notre feuille de route vise à guider les décideurs politiques dans la défense de la démocratie.

DÉFENDRE LA DÉMOCRATIE À LA FRONTIÈRE DE L'INTELLIGENCE ARTIFICIELLE

PAR MICHAEL BĄK

Directeur exécutif, Forum de l'Information et de la démocratie

La trajectoire de l'évolution technologique, y compris celle de l'intelligence artificielle, n'est pas inévitable, malgré les déclarations prophétiques des milliardaires et des cadres dirigeants de la Tech. Elle ne doit pas non plus être construite pour répondre aux priorités des intérêts privés, du capital et des actionnaires. L'intelligence artificielle est un bien public. Elle doit nous servir non seulement en tant que consommateurs, mais surtout en tant que citoyens.

Nos institutions démocratiques ont la responsabilité de façonner et d'orienter l'évolution de l'IA dans une direction qui soit conforme aux valeurs communes de nos démocraties, qui respectent le libre arbitre des individus partout dans le monde, et qui renforce nos droits humains fondamentaux, y compris le droit à une information fiable.

Lancé en 2019 et comprenant à ce jour 52 États signataires, le Partenariat international pour l'information et la démocratie vise à garantir que nos institutions démocratiques gouvernent l'espace global de l'information et de la communication selon des règles et des normes démocratiques, élaborées en partenariat avec la société civile. Notre organisation, le Forum sur l'information et la démocratie, est l'entité dirigée par la société civile qui donne vie à cette vision. Nous nous concentrons sur la prévention, la limitation et l'atténuation des pressions exercées sur nos démocraties par des technologies non réglementées, y compris l'intelligence artificielle.

L'intelligence artificielle présente une transformation sans précédent dans la manière dont nous créons, diffusons et consommons l'information. Elle décide ce que vous voyez et ce que je vois ; et nous ne voyons pas les mêmes choses. Ces systèmes permettent à n'importe qui de créer et de diffuser facilement de l'information, mais ils sont souvent biaisés, discriminatoires à l'égard de certains groupes, ou hallucinatoires. Les systèmes d'IA peuvent également être facilement détournés par des acteurs malveillants qui cherchent à tromper les citoyens, à influencer les processus politiques et à semer le doute sur les faits qui constituent le fondement du discours démocratique.

Nos recommandations visent à anticiper, à prévenir ces préjudices et à orienter l'innovation technologique en tant que bien public, dans une direction qui sert l'intérêt public.

Nous ne devons pas commettre les mêmes erreurs qu'auparavant, lorsque les réseaux sociaux et les entreprises technologiques décidaient des règles du jeu, fixaient les agendas, déterminaient quels préjudices importaient (et où), et s'emparaient des récits politiques. Cela a entraîné de trop nombreuses excuses de la part des entreprises et, malheureusement, des dommages bien substantiels à nos communautés, à nos institutions démocratiques et à nos choix en tant que citoyens.

Nous pensons que l'intelligence artificielle peut emprunter une voie différente, plus éclairée et plus significative. Cette technologie peut progresser sur une voie guidée par le contrôle démocratique, constamment évaluée et améliorée grâce au leadership civique et à la participation inclusive.

Nous devons garantir des cadres et des mécanismes inclusifs permettant aux citoyens de s'assurer que les systèmes d'IA sont développés et déployés dans l'intérêt de notre monde diversifié. Et cela ne peut se faire qu'avec transparence, responsabilité et contrôle démocratique.

Ce rapport présente des recommandations pour y parvenir. Et cela peut se faire tout en encourageant une innovation florissante.

Le 28 septembre 2023, à l'occasion de la Journée internationale de l'accès universel à l'information, nous avons réuni quatorze experts de premier plan de diverses régions et disciplines pour entamer un processus inclusif et consultatif visant à élaborer des recommandations politiques d'une importance cruciale pour les gouvernements et les entreprises d'IA. Sous la coprésidence de Laura Schertel Mendes, avocate et professeure de droit civil à l'IDP au Brésil, et de Jonathan Stray, chercheur principal au Berkeley Center for Human-Compatible AI aux États-Unis, le groupe représente un pôle de réflexion significatif. Près de la moitié du groupe provient de « la majorité mondiale », la moitié sont des femmes, et tous représentent une grande diversité d'expériences vécues, de rigueur académique et d'engagement en faveur d'une IA responsable et éthique.

L'équipe de rapporteurs, dirigée par Viviana Padelli, économiste et experte en politique publique, et Katharina Zügel, Policy Manager au Forum sur l'information et la démocratie, a interrogé et consulté plus de 150 personnes issues de divers milieux à travers le monde dans le cadre de ce processus. Sous l'aide du groupe de travail, elles ont formulé des actions urgentes pour les décideurs mondiaux, les entreprises d'IA et la société civile alors que nous avançons tous vers un avenir façonné par l'IA.

Ce rapport couvre une large gamme de considérations et de mesures actives que les décideurs politiques et les entreprises peuvent prendre pour garantir que l'IA serve nos démocraties et nos citoyens.

Nous sommes profondément reconnaissants pour les contributions de l'ensemble du Groupe de travail et de nos partenaires à travers le monde qui ont déployé des efforts considérables pour que nous réussissions à mettre l'IA au service de nos démocraties, de nos concitoyens et des générations futures.

Le temps presse. Aucune technologie dans l'histoire de l'humanité ne s'est développée à une vitesse aussi exponentielle. Et nous devons suivre le rythme.

RECOMMANDATIONS PRINCIPALES

Les entreprises et entités spécialisées dans l'IA peuvent concevoir de manière proactive des systèmes d'IA plus inclusifs et fiables par les moyens suivants :

- Mise en place d'un **processus inclusif et participatif** prévoyant au minimum une participation équitable, durable et substantiel de chercheurs indépendants, de la société civile et des communautés concernées, afin de définir **les règles et les critères régissant la provenance et la curation des données, l'étiquetage humain pour l'apprentissage de l'IA, leur harmonisation et le red-teaming**. Ces règles et critères devraient être accessibles au public.
- Mise en œuvre de **mesures d'atténuation des risques et de modération des résultats**, y compris sous forme de **mécanismes de notification et d'action** pour recueillir les réactions des utilisateurs, de **dispositifs de réclamation**, de collaboration avec **des signaleurs de confiance et des fact-checkers**, ainsi que d'initiatives de **red-teaming** en continu pour répondre aux risques identifiés.
- Possibilité pour les utilisateurs de choisir facilement et intuitivement des **systèmes de recommandation alternatifs** ne s'appuyant pas sur l'optimisation de l'engagement, mais sur un référencement en faveur de résultats individuels et sociétaux positifs, tels que des informations fiables, des contenus fédérateurs ou des informations diversifiées.
- Mise en œuvre d'une politique prévoyant que le contenu et les utilisateurs doivent obtenir un « **droit de recommandation** » avant de faire l'objet d'une promotion ou de figurer dans les fils d'actualité. Ce droit devrait être accordé en vertu d'une signature cryptographique valide associée à des entités de confiance.
- **La communication claire,** facilement accessible et visible, **concernant les données et les interactions des utilisateurs avec un système d'IA, de toute activité de traitement des données, de la manière dont les données sont stockées et du partage éventuel de ces données avec des tiers.** Les utilisateurs devraient pouvoir **refuser** que leurs **données entrantes et leurs interactions soient stockées et utilisées,** et l'opt-out (la dérogation) devrait être **le paramètre par défaut**.

Une réglementation est nécessaire pour contraindre les entreprises et les entités spécialisées dans l'IA à mettre en place des systèmes d'IA responsables, en leur imposant les obligations suivantes :

- Fournir des **informations sur leurs jeux de données d'apprentissage à des fins d'examen public**, de manière facilement accessible et compréhensible incluant une base de données consultable.
- Effectuer des **analyses d'impact pour vérifier la présence de biais, y compris en matière de diversité et de représentation, ainsi que les inexactitudes et les fausses représentations dans différentes langues**, *avant* le déploiement des systèmes d'IA, et les réexaminer en continu.
- **Mettre en place des structures de gouvernance démocratique,** pouvant prendre différentes formes, telles qu'un **conseil de surveillance, une assemblée de citoyens ou une structure de représentation des employeurs et des utilisateurs.**
- Procéder à une évaluation des risques systémiques, en examinant les risques pour l'espace informationnel avant le déploiement, et procéder à une évaluation de la conformité par un tiers pour les systèmes à risque moyen ou élevé.

Des mesures sont nécessaires pour renforcer la confiance dans l'espace informationnel, telles que :

- **Définir des normes régissant l'authenticité et la provenance des contenus, y compris pour authentifier des auteurs,** et appliquer ces normes dans les communications gouvernementales et les médias.
- **Obliger les plateformes à identifier les informations relatives à l'authenticité et à la provenance et les contenus générés par l'IA** par les meilleurs moyens actuellement disponibles et à **présenter ces informations à l'utilisateur final.**
- **Obliger les déployeurs d'entités synthétiques** (p. ex. chatbots, assistants virtuels) **à informer les utilisateurs qu'ils communiquent avec un système interactif piloté par l'IA et à prévoir des méthodes permettant de détecter de manière fiable le contenu qu'ils génèrent, notamment en intégrant un filigrane dans le contenu généré.**
- Envisager l'octroi d'un financement public pour soutenir le développement et la maintenance d'une **infrastructure publique pour des systèmes d'IA dignes de confiance**. Celle-ci comprendrait **des jeux de données publiques, des alternatives publiques aux systèmes de recommandation, de modération de contenu ou de référencement à but lucratif, des alternatives publiques aux systèmes d'IA générative à but lucratif et des infrastructures alternatives pour l'espace informationnel numérique.**
- Promouvoir la création d'un **système de certification sur mesure pour les entreprises d'IA**, sur le modèle du **système de certification du commerce équitable**.
- **Adopter des codes et des lignes directrices pour encourager une utilisation responsable des systèmes d'IA par les gouvernements et les médias,** tels que la charte de Paris sur l'IA et le journalisme, afin de définir des lignes directrices claires sur l'utilisation de l'IA.
- Instaurer des **droits pour les médias et les journalistes, notamment le droit de savoir** (concernant l'utilisation de leur contenu dans les systèmes d'IA), **le droit de retrait (*opt-out*) et le droit à une compensation équitable**.

Établir des responsabilités claires pour les préjudices subis dans l'espace informationnel suppose de :

- Mettre en place **un régime de responsabilité fondé sur la faute pour les développeurs et les déployeurs d'IA** concernant les résultats de leurs systèmes. Les développeurs et les déployeurs seraient tenus pour responsables de leur manquement aux obligations liées aux mesures d'atténuation des risques, aux exigences de transparence et au devoir de diligence, sauf preuve du contraire. En outre, il devrait être clairement établi que **la charge de la preuve est du ressort des développeurs et des déployeurs d'IA** dans les cas où des individus ou un groupe auraient subi des préjudices.
- Instaurer un **régime de responsabilité stricte pour les développeurs et les déployeurs de systèmes d'IA utilisés pour le micro-ciblage des utilisateurs en fonction de caractéristiques protégées ou de catégories particulières de données à caractère personnel.**
- **Instaurer une présomption réfutable de responsabilité des plateformes pour les contenus illégaux qu'elles hébergent et les préjudices qu'elles causent, à moins qu'elles ne puissent prouver qu'elles ont** mis en place des mesures globales d'atténuation des risques, respecté les exigences de transparence et adopté des normes de pointe en matière de détection, d'étiquetage, de provenance et d'authenticité.

- Clarifier juridiquement le fait que **le contenu généré par l'IA** ne doit **pas être considéré comme un contenu de tiers ou un contenu d'hébergement** par le système d'IA pour déterminer la responsabilité du déployeur de l'IA générative.
- **Définir un cadre juridique exhaustif précisant expressément les droits des personnes,** notamment le droit d'être informé, de recevoir une explication, de contester un résultat et de ne pas faire l'objet d'une quelconque discrimination**, et obliger les systèmes d'IA à mettre en place des procédures de traitement des réclamations**.
- **Nommer un ombudsman de l'IA ou renforcer toute institution de médiation existante** chargée d'examiner les réclamations non réglées et de servir de représentant du plaignant si aucune solution à l'amiable ne peut être trouvée.

Les systèmes d'IA doivent faire l'objet d'une supervision et d'un contrôle indépendants :

- **Créer une nouvelle autorité (ou renforcer les capacités des autorités existantes)** et lui attribuer un mandat de supervision de l'application des réglementations en matière d'IA et de publication des actes exécutoires.
- **Créer et financer un organisme indépendant de recherche sur l'IA,** national ou supranational, composé de plusieurs laboratoires de recherche indépendants.
- **S'engager à indemniser financièrement les OSC qui participent aux institutions et structures officielles de contrôle et de gouvernance de l'IA.**
- Obliger les développeurs et déployeurs d'IA à garantir la transparence de leurs systèmes dans le cadre d'une approche graduelle, en fournissant des informations au grand public et des informations plus détaillées aux régulateurs et aux chercheurs agréés.
- Contraindre les plateformes à accorder à des chercheurs agréés la possibilité de mener des évaluations expérimentales de systèmes d'IA et de mettre en place un « bac à sable de responsabilisation » accessible aux partenaires extérieurs.
- **Instaurer une taxe sur les entreprises et les entités du secteur de l'IA afin de tenir compte de l'impact sociétal de l'IA.** Une partie des recettes générées par cette taxe devrait être affectée au financement de programmes communautaires de l'éducation à l'IA, d'alternatives publiques aux systèmes à but lucratif et d'initiatives de la société civile.
- **Mettre en place des protections juridiques rigoureuses pour les lanceurs d'alerte** anciens ou actuels employés de l'industrie de l'IA.
- Garantir que la gouvernance internationale de l'IA soit régie **par des principes démocratiques en renforçant la coopération par le biais du Partenariat pour l'information et la démocratie, et en** promouvant la formation d'un **Forum mondial de l'IA pour un dialogue ouvert, impliquant une participation régulière et égale de la société civile, des médias et des journalistes, des chercheurs, et d'autres organisations communautaires et d'intérêt public.**

LE **GROUPE DE TRAVAIL**

Le groupe de travail sur l'intelligence artificielle et ses implications pour l'espace de l'information et de la communication a été constitué le 28 septembre 2023, à l'occasion de la Journée internationale de l'accès universel à l'information. Ce groupe de travail est composé de 14 chercheurs et experts de renommée mondiale issus de diverses disciplines universitaires et appliquées, qui ont guidé les rapporteurs et le Forum dans la définition de l'axe thématique du rapport et l'élaboration de ses recommandations.

COPRÉSIDENTS DU GROUPE DE TRAVAIL :

- **Laura Schertel Mendes,** professeur de droit à l'Institut brésilien pour le développement, l'éducation et la recherche (IDP), et rapporteuse de la commission de juristes chargée de conseiller le Sénat brésilien sur la réglementation de l'IA, Brésil

- **Jonathan Stray,** chercheur émérite, UC Berkeley Center for Human-Compatible AI, États-Unis

MEMBRES DU GROUPE DE TRAVAIL :

- **Rachel Adams,** directrice de l'Indice mondial de l'IA responsable et de l'Observatoire africain de l'IA responsable, Research ICT Africa, Afrique du Sud

- **Linda Bonyo,** directrice fondatrice, Africa Law Tech ; fondatrice du Lawyers Hub, Kenya

- **Marta Cantero Gamito,** professeure de droit des technologies de l'information, université de Tartu ; chargée de recherche, Florence School of Transnational Governance (chaire sur l'IA et la démocratie, EUI), Italie

- **Alistair Knott, professeur d'IA,** Université Victoria de Wellington, Nouvelle-Zélande

- **Syed Nazakat,** fondateur et directeur général, DataLEADS, Inde

- **Alice Oh,** professeure, Korea Advanced Institute of Science and Technology, Corée

- **Prof. Alejandro Pisanty,** Faculté de chimie, Université nationale autonome du Mexique (UNAM), Mexique

- **Gabriela Ramos,** Sous-Directrice générale pour les Sciences Sociales et Humaines, UNESCO

- **Prof. Dr Achim Rettinger,** linguistique informatique, université de Trèves, Allemagne

- **Prof. Edward Santow,** codirecteur, Institut de technologie humaine, Université de technologie de Sydney, Australie

- **Dr Suzanne Vergnolle,** maître de conférences en droit des technologies, Cnam, France

- **Claes de Vreese,** professeur universitaire distingué d'IA et de société, spécialisé dans les médias et la démocratie, Université d'Amsterdam, Pays-Bas

RAPPORTEUSE PRINCIPALE :

- **Viviana Padelli,** économiste et professionnelle des politiques publiques, travaillant au croisement de la technologie et de la démocratie

RAPPORTEUSES :

- **Lia Chkhetiani,** politologue et avocate
- **Kaye Celine Palisoc,** analyste en informatique et en politiques publiques

FORUM SUR L'INFORMATION ET LA DÉMOCRATIE :

- **Christophe Deloire,** président
- **Michael Bąk,** directeur exécutif
- **Camille Grenier,** directeur des opérations
- **Katharina Zügel,** policy manager
- **Julie Pailhès,** chargée de projet

Le rapport de ce groupe de travail reflète les points de vue exprimés lors des discussions des équipes de rapporteurs avec les membres, des entretiens avec les spécialistes et des contributions écrites reçues de la part d'experts et d'organisations actives dans ce domaine. L'équipe de rapporteurs n'a pas recherché l'unanimité sur chaque conclusion ou recommandation, reconnaissant qu'il n'était pas toujours possible de concilier des points de vue différents. Ce rapport ne doit pas être considéré comme le résultat d'une négociation formelle validée par les membres de ce groupe de travail, mais comme le fruit de tous les efforts fournis par l'équipe de rapporteurs pour proposer une marche à suivre.

A PROPOS **DU FORUM SUR L'INFORMATION ET LA DÉMOCRATIE**

Pour des garanties démocratiques dans l'espace global de la communication et de l'information

Commission sur l'information et la démocratie

Composition :
Lauréats du Nobel (paix, économie, littérature), spécialistes des nouvelles technologies, journalistes.

Mission :
- A publié la Déclaration internationale qui fixe les principes universels pour l'information et la démocratie et inspire le Partenariat

PARTENARIAT INTERNATIONAL POUR L'INFORMATION ET LA DÉMOCRATIE

Composition :

États s'engageant sur des principes démocratiques dans l'espace de l'information et de la communication

Sommets annuels pour l'information et la démocratie

Objectifs :
- Renforcer la coopération internationale
- Travailler à la mise en œuvre des recommandations du Forum
- Favoriser la discussion entre société civile et gouvernements

ÉVALUATIONS

RECOMMANDATIONS

UN MÉCANISME DE MISE EN ŒUVRE

PUBLIE un rapport d'évaluation en amont des sommets du Partenariat

ÉVALUE les normes, structures et architectures de choix de l'espace de l'information et de la communication

DÉVELOPPE des cadres pour la régulation à destination des États

RASSEMBLE l'état de l'art de la recherche et des recommandations

Forum Information & Démocratie

Fondé par 11 organisations de la société civile et du monde de la recherche

Missions :
- Évalue l'espace de l'information et de la communication
- Produit des recommandations
- Accompagne des projets

HÉBERGE

ORGANISE

MOBILISE

Observatoire international sur l'information et la démocratie

Missions :
- Fournit aux États et à la société dans son ensemble des évaluations périodiques sur l'espace de l'information et de la communication

Groupes de travail internationaux

Missions :
- Rassemble des contributions internationales
- Propose des recommandations aux États

Organisations de la société civile et monde de la recherche

Missions :
- Contribuent au développement des recommandations et à l'évaluation de l'espace informationnel
- Promeuvent la mise en œuvre de garanties démocratiques dans cet espace

DÉFINITIONS

Agent d'intelligence artificielle : système d'intelligence artificielle caractérisé par sa capacité à exécuter de manière indépendante des actions contribuant de manière cohérente à la réalisation d'objectifs complexes définis par l'homme sur de longues périodes, et sous une supervision humaine directe minimale[1].

Algorithme : processus de calcul utilisé pour prendre des décisions[2].

Alternative publique au système à but lucratif : système développé et géré dans l'intérêt public par une ou plusieurs organisations agissant dans l'intérêt public. Cet organe peut prendre diverses formes telles qu'une organisation de recherche ou de la société civile, une autorité administrative indépendante, ou avoir un statut similaire à celui d'un média de service public, entre autres.

API (interface de programmation d'applications) : dispositif permettant à deux composants logiciels de communiquer entre eux à l'aide d'un ensemble de définitions et de protocoles[3].

Augmentation des données : technique consistant à accroître artificiellement la quantité de données d'entraînement en créant des copies modifiées d'un ensemble de données à l'aide de données existantes. Elle comprend des modifications mineures de l'ensemble de données ou l'utilisation du deep learning pour générer de nouveaux points de données[4].

Bac à sable (sandbox) de responsabilisation : environnement de test qui permet aux parties prenantes externes de saisir des données afin de mieux comprendre le fonctionnement d'un système d'intelligence artificielle.

Bot (bots des réseaux sociaux) : programmes dont la taille varie en fonction de leur fonction, de leur capacité et de leur conception. Ils peuvent être utilisés sur les plateformes de réseaux sociaux pour effectuer diverses tâches utiles ou malveillantes en simulant des comportements humains[5].

Cartes de modèles : documentation sur les utilisations et les limites de l'IA, les stratégies d'atténuation et les descriptions des méthodes de formation et d'essai prévues - avant et durant le développement du système.

Classificateurs : algorithme qui ordonne ou classe automatiquement les données en une ou plusieurs « classes ». Le processus de catégorisation ou de classification des informations en fonction de certaines caractéristiques est désigné par le terme de classification[6].

Curation : décisions liées au filtrage et au classement du contenu à l'attention de l'homme[7].

Curation des données : processus d'organisation, de description, de nettoyage, d'amélioration et de préservation des données pour un usage public[8].

Désinformation : informations fausses et délibérément créées pour nuire à une personne, un groupe social, une organisation ou un pays[9].

---

1 Shavit, Y. et al. (2023). *Practices for Governing Agentic AI Systems.* Disponible sur : https://openai.com/research/practices-for-governing-agentic-ai-systems (Consulté le 5 février 2024)

2 Tel que défini par le Forum sur l'information et la démocratie (2023). *Pluralisme de l'Information dans les Algorithmes d'Indexation et de Curation*, p. 14. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Pluralism_FR.pdf (Consulté le 8 février 2024).

3 Amazon Web Services (2023). *What is an API? - API Beginner's Guide - AWS*. Amazon Web Services, Inc. Disponible sur : https://aws.amazon.com/what-is/api/ (Consulté le 2 février 2024).

4 Awan, A.A. (2022). *A Complete Guide to Data Augmentation*. Disponible sur : www.datacamp.com/tutorial/complete-guide-data-augmentation (Consulté le 31 janvier 2024).

5 Tel que défini par la US Office of Cyber and Infrastructure Analysis (2018). Social Media Bots Overview. Disponible sur : https://niccs.cisa.gov/sites/default/files/documents/pdf/ncsam_socialmediabotsoverview_508.pdf?trackDocs=ncsam_socialmediabotsoverview_508.pdf (Consulté le 8 février 2024).

6 DeepAI. (2019). *Classifier*. Disponible sur : https://deepai.org/machine-learning-glossary-and-terms/classifier (Consulté le 2 février 2024).

7 Tel que défini par le Forum sur l'information et la démocratie (2023). *Pluralisme de l'Information dans les Algorithmes d'Indexation et de Curation et* , p 14. Disponible sur : https://informationdemocracy.org/pluralism/ (Consulté le 8 février 2024).

8 ICPSR. *Data Management & Curation*. Disponible sur : www.icpsr.umich.edu/web/pages/datamanagement/index.html (Consulté le 8 février 2024).

9 Tel que défini par l'UNESCO (2020). Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training. Disponible sur : https://en.unesco.org/fightfakenews (Consulté le 8 février 2024).

Déployeur d'IA : détermine l'utilisation du système d'IA, assure le contrôle du risque associé et tire profit de son fonctionnement[10].

Développeur IA : conçoit, code ou produit des systèmes d'IA[11].

Données d'apprentissage : informations/données employées au cours du processus d'apprentissage du modèle.

Données synthétiques : informations générées par informatique pour accroître ou remplacer les données réelles afin d'améliorer les modèles d'IA, de protéger les données sensibles et d'atténuer les biais[12].

Entités synthétiques : constructions artificielles, créées sous forme numérique ou physique, conçues pour imiter ou reproduire des caractéristiques spécifiques d'entités naturelles (p. ex. chatbots, assistants virtuels)[13].

Entreprises et entités d'IA : entreprises et tout autre acteur (p. ex. organisations à but non lucratif, instituts de recherche, etc.) qui développent et/ou mettent en œuvre des systèmes d'intelligence artificielle. Le développement et/ou déploiement de l'IA peut ne constituer qu'une partie de leur activité.

Étiquetage humain pour la formation de l'IA : participation humaine au processus d'étiquetage des données de formation[14].

Évaluation de la conformité : forme de contrôle humain *ex ante* visant à vérifier que les systèmes d'IA respectent les normes techniques, éthiques et juridiques reconnues[15].

Fine-tuning : procédure d'ajustement des paramètres d'un modèle linguistique préformé conséquent pour une tâche ou un domaine spécifique. Le fine-tuning permet de remédier au manque de spécialisation du modèle dans certains domaines en lui offrant la possibilité d'apprendre à partir de données spécifiques à un domaine afin d'améliorer sa précision et son efficacité pour des applications ciblées[16].

Grands modèles de langage (LLM) : un type de modèle d'intelligence artificielle ayant été formé par des algorithmes de deep learning pour reconnaître, générer, traduire et/ou résumer de grandes quantités de langage humain écrit et de données textuelles[17].

Harmonisation : opération consistant à amener les systèmes d'IA à se comporter conformément aux intentions et aux valeurs humaines[18].

IA générative : l'IA générative désigne une catégorie d'algorithmes d'IA qui génèrent de nouveaux résultats à partir des données sur lesquelles ils ont été formés[19].

Intelligence artificielle : système fonctionnant grâce à une machine capable d'influencer son environnement en produisant des résultats (tels que des prédictions, des recommandations ou des décisions) pour répondre à un ensemble donné d'objectifs définis par l'homme[20].

---

10 Law Insider. (n.d.). *deployer Definition*. Disponible sur : www.lawinsider.com/dictionary/deployer (Consulté le 2 février 2024).
11 BSA (2023). *AI Developers and Deployers: An Important Distinction. The Software Alliance*. Disponible sur : www.bsa.org/policy-filings/ai-developers-and-deployers-an-important-distinction (Consulté le 2 février 2024).
12 Martineau, K. (2021). *What is synthetic data?* [online] IBM Research Blog. Disponible sur : https://research.ibm.com/blog/what-is-synthetic-data (Consulté le 31 janvier 2024).
13 Nanni, D. (2023). *Synthetic Entities: Definitions, Characteristics, and Future Perspectives*. Brass For Brain. Disponible sur : https://medium.com/brass-for-brain/synthetic-entities-definitions-characteristics-and-future-perspectives-49673f22f6fe (Consulté le 31 janvier 2024).
14 Mehta, R. (2023). *Human Data Labeling for Successful AI*. iMerit. Disponible sur : https://imerit.net/blog/human-data-labeling-for-successful-ai/ (Consulté le 9 février 2024).
15 Mökander, J., Axente, M., Casolari, F. and Floridi, L. (2021). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, 32. Disponible sur : https://doi.org/10.1007/s11023-021-09577-4 (Consulté le 7 février 2024).
16 Turing (n.d.). *Fine-Tuning LLMs: Overview, Methods & Best Practices*. Disponible sur : www.turing.com/resources/finetuning-large-language-models#what-is-fine-tuning (Consulté le 9 février 2024).
17 Sarno, I. (2023). *What Is a Large Language Model?* knowledge-centre-interpretation.education.ec.europa.eu. Disponible sur : https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model (Consulté le 8 février 2024).
18 Ji, J., et al. (2023). *AI Alignment: A Comprehensive Survey*. arXiv (Cornell University). Disponible sur : https://doi.org/10.48550/arxiv.2310.19852. (Consulté le 2 février 2024)
19 Routley, N. (2023). *What is generative AI? An AI explains*. World Economic Forum. Disponible sur : www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/ (Consulté le 9 février 2024).
20 Tel que défini par l'OCDE dans *OECD AI Principles overview*. Disponible sur : https://oecd.ai/en/ai-principles (Consulté le 7 février 2024).

Interopérabilité : capacité à transférer et à restituer des données utiles et d'autres informations entre différents systèmes, applications ou composants, y compris les plateformes[21].

Médias : organisations responsables de la création périodique d'information et de contenu et à sa diffusion pour laquelle il y a une responsabilité éditoriale, indépendamment des moyens et de la technologie utilisés pour la délivrer, qui est destinée à la réception par une proportion significative du public dans son ensemble, et qui peut avoir un impact notable sur ce dernier[22].

Mésinformation : information fausse mais qui n'a pas été créée dans l'intention de nuire[23].

Modèle d'IA : programme formé sur un ensemble de données afin d'identifier certains modèles ou prendre certaines décisions sans autre forme d'intervention humaine. Les modèles d'IA utilisent différents algorithmes pour traiter les données pertinentes et réaliser les tâches, ou résultats, pour lesquelles ils ont été programmés[24].

Modération : lorsque les plateformes numériques analysent et filtrent le contenu généré par les utilisateurs sur la base de leurs propres règles et directives afin d'apprécier s'il faut héberger ou continuer à héberger certains éléments de contenu en vertu de leurs conditions de service. Ces décisions incluent le retrait de contenus, de manière permanente, temporaire ou par zone géographique[25].

Octroi de licences (Licensing) : outil *ex ante* de contrôle humain, qui implique une évaluation complète du système d'IA au regard d'exigences juridiques, éthiques et techniques, sans laquelle les systèmes d'IA ne peuvent recevoir d'autorisation de déploiement[26].

Open Source : logiciel publié sous une licence open source qui offre la liberté de l'utiliser, de l'étudier, de le partager et de l'améliorer[27].

Organisation de la société civile : tout collectif de citoyens bénévoles à but non lucratif, organisé au niveau local, national ou international[28].

Plateforme : entités contribuant à structurer l'espace de l'information et de la communication par la création de moyens techniques, d'architecture et de normes d'information et de communication[29].

Provenance des données : documentation sur l'origine d'un élément de données et la manière dont il est parvenu à son état actuel[30].

Provenance et authenticité du contenu : informations relatives à l'origine et à l'histoire d'un élément de contenu numérique (image, vidéo, enregistrement audio, document)[31].

Red Teaming : processus d'émulation des capacités d'attaque ou d'exploitation d'un adversaire potentiel contre le dispositif de sécurité d'une entreprise, réalisé par un groupe d'utilisateurs agréés appelé « red team »[32].

---

21 Tel que défini par le Forum sur l'information et la démocratie (2023). *Pluralisme de l'Information dans les Algorithmes d'Indexation et de Curation*, p.14. Disponible sur : https://informationdemocracy.org/pluralism/ (Consulté le 8 février 2024).

22 Tel que défini par le Forum sur l'information et la démocratie (2023). *Pluralisme de l'Information dans les Algorithmes d'Indexation et de Curation*, p. 14. Disponible sur : https://informationdemocracy.org/fr/pluralisme/ (Consulté le 8 février 2024).

23 Tel que défini par le Forum sur l'information et la démocratie (2020). Pour mettre fin aux *Infodémies*, p. 16. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Infodemies_FR.pdf (Consulté le 7 février 2024).

24 IBM (n.d.). *What is an AI model? IBM*. Disponible sur : www.ibm.com/topics/ai-model (Consulté le 2 février 2024).

25 Tel que défini par le Forum sur l'information et la démocratie (2022). *Accountability Regimes for Social Networks and their Users*, p.10. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Responsabilite-reseaux-sociaux_FR.pdf (Consulté le 8 février 2024).

26 Malgieri, G. and Pasquale, F. (2024). Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review*,] 52, p.105899. Disponible sur : https://doi.org/10.1016/j.clsr.2023.105899.

27 FSFE - Free Software Foundation Europe). *What is Free Software*. Disponible sur : https://fsfe.org/freesoftware/freesoftware.en.html (Consulté le 7 février 2024).

28 Organisation des Nations Unies . *Civil society*. United Nations. Disponible sur : www.un.org/en/civil-society/page/about-us (Consulté le 8 février 2024).

29 Pour une explication plus détaillée, voir le Forum sur l'information et la démocratie (2020). Pour mettre fin aux *Infodémies*, p. 16. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Infodemies_FR.pdf (Consulté le 8 février 2024).

30 faculty.washington.edu. (n.d.). *About Data Provenance*. [online] Disponible sur : https://faculty.washington.edu/hazeline/ProvEco/generic.html (Consulté le 1er février 2024).

31 c2pa.org.) *FAQ - C2PA*.Disponible sur : https://c2pa.org/faq/ (Consulté le 31 janvier 2024)

32 NIST . *Computer Security Resource Center - Glossary*. Disponible sur : https://csrc.nist.gov/glossary/term/Red_Team (Consulté le 8 février 2024).

Responsabilité fondée sur la faute : en droit civil, la responsabilité fondée sur la faute est une notion juridique selon laquelle un défendeur n'est tenu responsable des conséquences de ses actes que si sa faute (intention ou négligence) a été prouvée. En pratique, ceci signifie que le plaignant doit prouver que le dommage a été causé par la faute du défendeur[33].

Responsabilité stricte : en droit civil, la responsabilité stricte est une norme juridique en vertu de laquelle un défendeur est tenu responsable des conséquences de ses actes, quelle que soit sa faute (intention ou négligence). En pratique, ce principe signifie que la faute n'est pas un facteur nécessaire à la détermination de la responsabilité[34].

Responsabilité pénale : norme juridique en vertu de laquelle les individus sont tenus responsables de leurs actes s'ils ont commis un acte criminel[35].

Segments de données : sous-ensembles d'un jeu de données, généralement rassemblés en fonction de caractéristiques similaires.

Sujet IA : toute entité (p. ex. personne, agence, organisation, etc.) concernée par un système d'IA (p. ex. dont les données sont utilisées dans l'apprentissage, au sujet de laquelle les résultats de l'IA sont générés).

Systèmes de recommandation : systèmes qui suggèrent algorithmiquement des contenus à un utilisateur, en se basant potentiellement sur des informations concernant cet utilisateur (profilage en fonction de ses intérêts) ; portant sur ce contenu (y compris des signaux tels que l'indexation et/ou la prédiction de la viralité) ; et/ou dépendant des intérêts des organisations[36].

Système d'IA : composant, logiciel et/ou matériel basé sur l'IA. Généralement, les systèmes d'IA sont intégrés en tant que composants de systèmes plus étendus, plutôt qu'en tant que systèmes autonomes[37].

Tatouage numérique (Watermarking) : technique permettant d'incorporer un motif d'identification dans un élément de contenu afin d'en retracer l'origine[38].

Test A/B : technique consistant à comparer les performances de deux versions d'un même contenu pour déterminer laquelle attire davantage les visiteurs/téléspectateurs. Le test compare une version de contrôle (A) à une version variante (B) afin de déterminer laquelle est la plus performante en fonction d'indicateurs clés[39].

Traitement du langage naturel : ensemble de méthodes permettant de rendre le langage humain accessible aux ordinateurs[40].

Troll : entité qui perturbe délibérément les communautés en ligne[41].

Utilisateur d'IA : personne ou entité utilisant des systèmes d'IA.

---

33 Coleman, J.L. (2002). Fault and strict liability. *Risks and Wrongs*, pp.212–233. doi: https://doi.org/10.1093/acprof:oso/9780199253616.003.0012 (Consulté le 8 février 2024).

34 J.L. (2002). Fault and strict liability. *Risks and Wrongs* , op. Cit Coleman, J.L. (2002). Fault and strict liability. *Risks and Wrongs*, pp.212–233. doi: https://doi.org/10.1093/acprof:oso/9780199253616.003.0012 (Consulté le 7 février 2024).

35 Sneha Solanki (2024). *What is criminal liability? Definition and resources for defense attorneys*. Thomson Reuters Law Blog. Disponible sur : https://legal.thomsonreuters.com/blog/what-is-criminal-liability/(Consulté le 9 février 2024).

36 Tel que défini par le Forum sur l'information et la démocratie (2023). *Pluralisme de l'Information dans les Algorithmes d'indexation et de Curation*, p. 14. Available at: https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Pluralism_FR.pdf (Accessed: 8 February 2024).

37 Tel que défini par la Commission Européenne (2018). *A Definition of AI: Main Capabilities and Scientific Disciplines*, p.1. Disponible sur : https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (Consulté le 8 février 2024).

38 Brookings. . *Detecting AI fingerprints: A guide to watermarking and beyond*. Disponible sur : www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/ (Consulté le 31 janvier 2024).

39 Oracle.com (2022). *What is A/B Testing?* Disponible sur : www.oracle.com/cx/marketing/what-is-ab-testing/ (Consulté le 2 février 2024).

40 Eisenstein, J. (2019). *Introduction to natural language processing*. Cambridge, Massachusetts: The Mit Press.

41 Schwartz, M. (2008). The Trolls Among Us. *The New York Times*. 3 Aug. Disponible sur : www.nytimes.com/2008/08/03/magazine/03trolls-t.html (Consulté le 7 février 2024).

INTRODUCTION

Les capacités de plus en plus avancées et la prolifération commerciale des systèmes d'intelligence artificielle (IA), tels que ChatGPT, les « smart bots » et la propagande informatique, illustrent leur omniprésence, leur sophistication et leur impact potentiel sans cesse plus profond sur les processus démocratiques, y compris dans l'espace de l'information et de la communication. Bien que les systèmes d'IA puissent également être exploités en faveur des processus démocratiques, ils posent des défis considérables et laissent de nombreuses questions en suspens, qu'il convient d'examiner avec la plus grande attention[42].

C'est la raison pour laquelle la réglementation du développement, du déploiement et de l'utilisation de l'IA constitue une priorité majeure pour les responsables politiques. Parmi les initiatives visant à réglementer les systèmes d'IA, figurent notamment le décret sur l'IA du président Joe Biden aux États-Unis[43], la signature de la déclaration de Bletchley par plus de 20 pays[44], les *Principes directeurs internationaux du processus d'Hiroshima pour les systèmes d'IA avancés* publiés par les pays du G7 en 2023[45], l'adoption par l'UNESCO des *Recommandations sur l'éthique de l'intelligence artificielle* en 2021[46] et par l'OCDE de ses *Principes sur l'intelligence artificielle* en 2019[47], les négociations en cours pour une convention-cadre du Conseil de l'Europe sur l'IA[48], une loi de l'UE sur l'intelligence artificielle[49], le projet de loi brésilien numéro 2338 sur l'utilisation de l'IA[50] et le projet de loi canadien sur les l'intelligence artificielle et les données[51].

La nécessité de répondre aux menaces que l'IA fait peser sur l'écosystème de l'information et de la communication est particulièrement impérieuse en 2024, alors que trois milliards de personnes dans le monde sont appelées à voter lors d'élections majeures[52]. Les implications que les tout récents systèmes d'IA générative peuvent avoir sur le processus démocratique restent inconnus, notamment en termes de micro-ciblage et de fabrication de personnifications de candidats visant à manipuler les électeurs[53].

---

42 Kreps, S. and Kriner, D. (2023). *How AI Threatens Democracy. Journal of Democracy*. Disponible sur : www.journalofdemocracy.org/articles/how-ai-threatens-democracy/ (Consulté le 7 février 2024).

43 The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* Disponible sur : www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Consulté le 7 février 2024).

44 *The Bletchley Declaration by Countries Attending the AI Safety Summit.* 1-2 November 2023. Disponible sur : www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 (Consulté le 7 février 2024).

45 Commission Européenne (2023). *Principes directeurs internationaux du processus Hiroshima pour les systèmes d'IA avancés*. Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/hiroshima-process-international-guiding-principles-advanced-ai-system (Consulté le 7 février 2024).

46 UNESCO (2021). *Recommandation sur l'éthique de l'intelligence artificielle.* Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137_fre (Consulté le 7 février 2024).

47 OCDE (2019). *Recommandation du Conseil sur l'intelligence artificielle.* Disponible sur : https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449 (Consulté le 7 février 2024).

48 Conseil de l'Europe (2023). *Projet de Convention-cadre sur l'Intelligence Artificielle, les Droits de l'Homme, la Démocratie et l'Etat de Droit.* Disponible sur : https://rm.coe.int/cai-2023-28-fr-projet-de-convention-cadre/1680ae19a1 (Consulté le 7 février 2024).

49 Le 2 février, le Conseil des pays de l'UE a approuvé la version finale de la loi sur l'IA de l'UE, dont le projet avait également fuité en ligne. Le 13 février, les commissions du marché intérieur et des libertés civiles, de la justice et des affaires intérieures (LIBE) du Parlement européen ont approuvé un accord provisoire sur la loi sur l'IA, ouvrant la voie au vote final de l'assemblée législative prévu pour avril 2024. Bertuzzi, L. (2024). *EU countries give crucial nod to first-of-a-kind Artificial Intelligence law.* Euractiv. European Council (2024). *Artificial Intelligence Act: Council and Parliament strike a deal on the first rules for AI in the world.* Available at: www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ (Consulté le 7 février 2024). Disponible sur : www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/ (Consulté le 7 février 2024). La dernière version du projet de loi sur l'IA de l'Union Européenne a été consultée le 7 février 2024. Disponible sur : https://drive.google.com/file/d/13qcPGQNFHTcfg4XxlybibFIDnkWTHBbu/view?_hsmi=292273120&*hsenc=p2ANqtz-83IaGl4OZccMJdsws8KrlMygORyZSiS1yD7lHUY8dbIDZzVdNvqPoBwq5n2V3GczVGWk1nWLJyuDlHeodJ3HJxNK4ia403C9b8JBLyBxh4CZ7iPHc* (Consulté le 7 février 2024).
Yun Chee, F (2024). EU lawmakers ratify political deal on artificial intelligence rules. Reuters. Disponible sur : https://www.reuters.com/technology/eu-lawmakers-back-political-deal-artificial-intelligence-rules-2024-02-13/ (Consulté le 13 février 2024).

50 Le Congrès National du Brésil (2023). *Bill 2338/2023 'Brazilian Artificial Intelligence Act'.* Disponible sur : https://mcusercontent.com/af97527c75cf28e5d17467eaa/files/248d109f-eeef-7496-4df1-12d29affb522/PL_23382023_Senado_ENG_VF.pdf (Consulté le 7 février 2024).

51 Gouvernement du Canada (2023). *Loi sur l'intelligence artificielle et les données*. Disponible sur : https://ised-isde.canada.ca/site/innover-meilleur-canada/fr/loi-lintelligence-artificielle-donnees (Consulté le 7 février 2024).

52 Hsu, T., Thompson, S.A. and Myers, S.L. (2024). *Elections and Disinformation Are Colliding Like Never Before in 2024.* The New York Times. 9 Jan. Disponible sur : www.nytimes.com/2024/01/09/business/media/election-disinformation-2024.html (Consulté le 7 février 2024).

53 *Ibid*

Le *Global Risks Report 2024*, récemment publié par le Forum économique mondial, reconnaît également l'urgence de la question. Il estime en effet que la mésinformation et la désinformation alimentées par l'IA et les réseaux sociaux constituent les plus grandes menaces pour le monde à court terme, en raison notamment de leur impact négatif sur la crédibilité des élections, sur la confiance dans l'information et sur la polarisation politique[54].

La résolution des problèmes les plus urgents liés à l'influence de l'IA sur l'espace de l'information et de la communication appelle des mesures de grande envergure, immédiates et multipartites, définies par des principes et des processus démocratiques. Bien que les États et les organisations internationales agissent à la hauteur de leurs moyens, la plupart de ces initiatives ne portent pas nécessairement sur l'espace de l'information et de la communication et énoncent plus souvent des principes généraux que des actions politiques concrètes.

Le Forum sur l'information et la démocratie privilégie une approche globale pour préserver les règles démocratiques dans l'espace de l'information et de la communication, en proposant des recommandations de politiques publiques réalisables et exhaustives pour garantir que l'IA ne soit pas un risque, mais au contraire un outil favorisant l'intégrité de l'information.

Les développeurs et les déployeurs de systèmes d'IA sont responsables au premier chef de la conception de systèmes d'IA dignes de confiance et au service de l'intérêt public. Dans cette optique, le chapitre 1 définit des garde-fous pour le développement et le déploiement de systèmes d'IA affectant l'espace de l'information et de la communication. Le rapport propose également des régimes de responsabilité afin de défendre les droits, tant individuels que collectifs[55]. À cette fin, le chapitre 2 définit des régimes de responsabilité pour les développeurs, les déployeurs et les utilisateurs au sein de l'écosystème de l'information et de la communication. Outre leur caractère réactif (c.-à-d. en imposant le respect des règles et en prévoyant des mécanismes de réparation), les régimes de responsabilité devraient également encourager des approches proactives dans le but de créer un environnement propice à une IA éthique. En conséquence, le chapitre 3 présente des systèmes d'incitation favorisant le développement, le déploiement et l'utilisation éthiques des systèmes d'IA. Enfin, en complément des cadres de gouvernance aux niveaux national et international, il est essentiel de garantir l'implication des différentes parties prenantes, un contrôle minutieux de la part des chercheurs et des processus solides pour la diffusion publique des systèmes d'IA afin de faire valoir les règles démocratiques en matière d'IA. À cet égard, le chapitre 4 propose des options de gouvernance holistiques et la mise en place de règles visant à réguler et à gouverner les systèmes d'IA.

IA et création d'informations

Les systèmes d'IA, et plus particulièrement les systèmes d'IA générative, révolutionnent la façon dont nous créons des informations sur différents supports, notamment le texte, l'audio, les images et la vidéo, ce qui présente à la fois des difficultés et des opportunités :

- Les systèmes d'IA générative sont actuellement conçus et utilisés d'une manière qui pourrait ne pas respecter les législations en matière de droits d'auteur, de protection des données ou de

54 The World Economic Forum (2024). The Global Risks Report 2024. Disponible sur : www.weforum.org/publications/global-risks-report-2024/ (Consulté le 7 février 2024).
Associated Press (2024). *AI-powered misinformation is the world's biggest short-term threat, Davos report says.* Disponible sur https://apnews.com/article/artificial-intelligence-davos-misinformation-disinformation-climate-change-106a1347ca9f987bf71da1f86a141968 (Consulté le 7 février 2024).

55 Le concept de protection des droits collectifs est très courant au Brésil. À cet égard, le projet de loi sur l'IA (N° 2338, 2023) inclut cet aspect à l'article 6 : La défense des intérêts et des droits prévus dans cette loi peut être exercée devant les organismes administratifs compétents, ainsi qu'en justice, individuellement ou collectivement, conformément aux dispositions de la législation pertinente concernant les instruments de protection individuelle, collective et diffuse.
Le Congrès National du Brésil (2023). Projet de loi 2338/2023 «Loi brésilienne sur l'intelligence artificielle». Disponible sur : https://mcusercontent.com/af97527c75cf28e5d17467eaa/files/248d109f-eeef-7496-4df1-12d29affb522/PL_23382023_Senado_ENG_VF.pdf (Consulté le 7 février 2024).

respect de la vie privée. Cela inclut l'utilisation de contenu protégé par des droits d'auteur sans mention appropriée ou la fuite involontaire d'informations privées ou sensibles à partir de leurs données d'apprentissage, pouvant représenter un problème pour la liberté d'expression et la protection de la vie privée[56].

- L'IA générative remet en question l'authenticité et la fiabilité du contenu produit. L'IA est en mesure d'inventer des sources et de générer de la mésinformation (hallucinations) et des « deepfakes ». Ces informations peuvent être utilisées de manière involontaire ou intentionnelle pour favoriser la mésinformation et la désinformation, semer le chaos, tromper, ébranler la confiance et mettre en péril le débat public. En outre, les systèmes d'IA générative réduisent les obstacles à la production de contenus fallacieux, permettant à une plus grande diversité de propagandistes de créer et de diffuser à grande échelle des éléments de désinformation et de propagande plus convaincants, plus diversifiés et mieux adaptés[57].
- Le contenu généré par l'IA, influencé par les données sur lesquelles les systèmes sont entraînés, révèle souvent des biais à l'encontre des communautés historiquement marginalisées et minoritaires. Pour éviter que les systèmes d'IA n'exacerbent les inégalités, diverses précautions doivent être adoptées au cours de leur développement. Ainsi, si les jeux de données d'apprentissage sont plus représentatifs que les réalités dominantes, l'IA est en capacité d'amplifier les voix des groupes sous-représentés.
- Si l'IA générative est susceptible d'améliorer le tri, l'organisation et la personnalisation des informations, son utilisation par les journalistes et les organismes de médias sans un contrôle humain adéquat peut conduire à la diffusion d'informations biaisées ou non vérifiées, ce qui finit par éroder la confiance du public dans les médias.

IA et diffusion d'informations

Les systèmes d'IA, y compris ceux destinés à la modération de contenu, à la recommandation, au référencement et à l'IA générative, tiennent une place importante dans la diffusion de l'information en ligne. Leur utilisation est assortie de sérieuses implications :

- L'IA générative peut être utilisée pour créer des bots sur les réseaux sociaux, intensifier la mésinformation et la désinformation et leur donner plus de crédibilité, contribuant ainsi à un véritable chaos informationnel. Cette technologie peut également être utilisée pour lancer des armées de « trolls » artificiels opérant sur les réseaux sociaux, ainsi que pour des activités d'astroturfing[58], qui multiplient de manière spectaculaire l'ampleur et la portée des campagnes de désinformation et de propagande[59].
- Les systèmes d'IA sont chargés de décisions capitales au sein de l'espace de l'information et de la communication, dans la mesure où la quantité d'informations disponibles et de contenus créés dépassent les capacités humaines de consommation, de tri, de modération et de vérification. Les systèmes de recommandation de l'IA déterminent les informations que nous voyons, ce qui modifie considérablement le paysage informationnel et menace de créer des chambres d'écho et des bulles de filtrage[60].

---

56 Allen, D. and Weyl, E.G. (2024). The Real Dangers of Generative AI. Journal of Democracy. Disponible sur : www.journalofdemocracy.org/articles/the-real-dangers-of-generative-ai/ (Consulté le 7 février February 2024).

57 Goldstein, J. et al (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Disponible sur : https://arxiv.org/pdf/2301.04246.pdf (Consulté le 7 février 2024).

58 «L'astroturfing est une forme d'activité numérique manufacturée, stratégique et dont le but est de tromper. Elle est initiée par des acteurs politiques sur Internet, qui imite l'activité ascendante des utilisateurs.»
Kovic, M et al. (2018). Digital astroturfing in politics: Definition, typology, and countermeasures. Studies in Communication Sciences, 18(1). Disponible sur : www.hope.uzh.ch/scoms/article/view/j.scoms.2018.01.005/991 (Consulté le 7 février 2024).

59 Zhou, J et al (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. Disponible sur : https://jiaweizhou.me/assets/chi23_ai_misinfo.pdf (Consulté le 7 février 2024).

60 Arguedas, A. et al (2022). Echo chambers, filter bubbles, and polarisation: a literature review. Reuters Institute for the Study of Journalism. Disponible sur : https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review (Consulté le 7 février 2024).

- Les systèmes de référencement et de modération de l'IA présentent des biais, leurs processus décisionnels manquent de transparence et ils sont souvent optimisés pour l'engagement sans tenir compte des risques sociopolitiques que peuvent comporter le contenu et les intervenants qu'ils mettent en avant ou analysent. Non seulement les systèmes d'IA modèrent les contenus en s'appuyant sur leurs propres définitions et jeux de données d'apprentissage secrètes, mais de plus, les politiques de modération, les statistiques de précision ou les rapports de transparence ne sont généralement pas publiés.
- L'influence des systèmes d'IA, et de ce fait de leurs développeurs et déployeurs, sur l'espace de l'information et de la communication ne cesse de croître, ce qui renforce le pouvoir des entreprises technologiques privées sur les processus démocratiques.
- Les systèmes d'IA peuvent également être utilisés à mauvais escient par les États. Les gouvernements peuvent ordonner aux entités d'IA d'utiliser ces systèmes pour supprimer les contenus jugés indésirables, renforçant ainsi la censure. En règle générale, les utilisateurs ont peu de contrôle ou d'influence lorsqu'ils interagissent avec les systèmes de recommandation et de modération de l'IA.
- Bien que les contenus générés par l'IA soient largement diffusés, les plateformes en ligne qui les hébergent ne disposent pas de politiques exhaustives en la matière. En outre, il est souvent impossible de déterminer avec certitude qui doit être tenu pour responsable des dommages causés par les contenus générés par l'IA et diffusés sur les plateformes de réseaux sociaux, ce qui augmente considérablement le risque pour les victimes qui cherchent à obtenir une indemnisation[61].
- L'exploitation numérique des informations personnelles et la publicité micro-ciblée peuvent être multipliées par des algorithmes sophistiqués de référencement et de modération de contenu, ainsi que par l'IA générative, qui réduit les efforts de personnalisation grâce à l' « inférence de la personnalité » à partir d'éléments textuels[62].
- Les utilisateurs ne sont pas en mesure de distinguer les informations dont ils sont automatiquement exclus ou auxquelles ils sont explicitement exposés. Des messages ouvertement personnalisés peuvent être utilisés pour instrumentaliser leurs craintes et manipuler leur comportement, au détriment de la démocratie[63].

IA et consommation d'informations

L'IA a la capacité de modifier la façon dont nous percevons les informations que nous consommons :

- Il devient de plus en plus difficile de distinguer les contenus générés par l'IA de ceux créés par l'homme, ce qui altère la confiance dans l'espace informationnel mondial. Selon certaines études, les éléments de mésinformation générés par l'IA sont plus difficiles à identifier car ils répondent à des critères de crédibilité, de transparence et d'exhaustivité, rendant ainsi la propagande ou la désinformation moins facile à détecter[64]. En outre, les filigranes peuvent être supprimés ou modifiés, compliquant ainsi davantage la capacité à déceler de manière fiable la nature synthétique du contenu. Même lorsque des mécanismes de détection sont utilisés pour vérifier le contenu, ils peuvent être erronés, de sorte qu'il n'existe actuellement aucune méthode fiable pour déterminer sans équivoque si un contenu est généré par l'IA[65].

61 Karathanasis, A.L et al (2022). Civil Liability for AI Systems: Comment on EU Commission's Proposals. MIAI. Disponible sur : https://ai-regulation.com/eu-commission-proposals-on-ai-civil-liability/ (Consulté le 7 février 2024).
62 Simchon, A. et al. (2024). The persuasive effects of political microtargeting in the age of generative AI. Disponible sur : https://academic.oup.com/pnasnexus/article/3/2/pgae035/7591134 (Consulté le 7 février 2024).
63 Wheeler, T. (2023). The three challenges of AI regulation. Brookings. Disponible sur : www.brookings.edu/articles/the-three-challenges-of-ai-regulation/ (Consulté le 7 février 2024).
64 Zhou, J et al (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. Disponible sur : https://jiaweizhou.me/assets/chi23_ai_misinfo.pdf (Consulté le 7 février 2024).
65 Leibowicz, C. (2023). Why watermarking AI-generated content won't guarantee trust online. MIT Technology Review. Disponible sur : www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/ (Consulté le 7 février 2024).

- Les systèmes d'IA sont en mesure de fournir aux utilisateurs des informations mieux adaptées à leurs centres d'intérêt. Cependant, cette hyperpersonnalisation peut également conduire à un paysage informationnel fragmenté et nuire à notre capacité de discerner les informations dignes de confiance. La personnalisation des informations ciblées limite l'accès à des informations pluralistes, exacerbe la polarisation politique et entrave la construction du débat public.

À la lumière des enjeux et des limites décrits ci-dessus, les chapitres suivants présentent des recommandations pratiques et stratégiques visant à aider les décideurs à définir des cadres exhaustifs, nuancés et démocratiques permettant d'exploiter le potentiel d'innovation des systèmes d'IA tout en traitant efficacement les risques qu'ils représentent pour l'espace de l'information et de la communication.

CHAPITRE 1 :
DÉVELOPPER ET DÉPLOYER DES SYSTÈMES D'IA SÛRS ET RESPONSABLES DANS L'ESPACE DE L'INFORMATION ET DE LA COMMUNICATION

INTRODUCTION

Les systèmes d'intelligence artificielle (IA) redéfinissent fondamentalement notre façon de créer, d'accéder et d'interagir avec l'information. Ils jouent un rôle clé dans le façonnement de l'opinion publique, en influençant le discours démocratique et en dessinant l'avenir du journalisme. Les algorithmes pilotés par l'IA jouent un rôle déterminant dans la modération et la curation des contenus, et sont essentiels à la publicité hyper-ciblée basée sur la surveillance. Il convient ici de souligner que le fait que des interfaces intuitives ont une incidence considérable sur la facilité et la rapidité avec lesquelles sont créés les textes, sons, images et données synthétiques produits par les utilisateurs, qu'ils soient bienveillants ou malveillants.

Le secteur de l'IA est aujourd'hui dominé par un petit nombre d'entreprises à but lucratif situées dans les pays du Nord. Néanmoins, l'influence considérable exercée par ces intérêts privés sur le domaine de l'information n'est pas suffisamment encadrée par nos institutions démocratiques, qui doivent impérativement mettre en place des protocoles de sécurité, des normes éthiques et des cadres réglementaires pour garantir que l'IA bénéficie à l'ensemble du public. Ces lacunes en matière de contrôle démocratique permettent aux entreprises de donner la priorité à la domination du marché, à la croissance et à la valeur actionnariale plutôt qu'à la sécurité, à l'éthique et aux répercussions sur nos institutions, nos sociétés et l'architecture de l'espace informationnel.

Un déploiement prématuré de l'IA pourrait avoir de nombreuses conséquences négatives sur les citoyens, sur nos sociétés et sur les institutions démocratiques, en modifiant notre façon de créer, de diffuser et de consommer l'information :

- *L'hyperpersonnalisation du contenu par des algorithmes sophistiqués, alimentée par l'exploitation des données de surveillance et la rapidité de l'amplification algorithmique, peut conduire à la production d'éléments de désinformation et de mésinformation plus efficaces et plus manipulatoires, entraînant l'instrumentalisation de l'opinion publique, la ségrégation des utilisateurs au sein de chambres d'écho médiatiques et l'érosion de la diversité et du pluralisme des sources d'information.*
- *Les discours de haine, illégaux ou légaux (mais néanmoins toxiques), favorisent l'incitation à la violence, renforcent la polarisation et nuisent aux interactions entre les groupes.*
- *L'essor de l'IA générative rend de plus en plus difficile, voire impossible, l'identification de contenus authentiques, réduisant ainsi la capacité des utilisateurs finaux à vérifier la provenance et l'exactitude des informations et, in fine, la confiance dans l'espace informationnel.*
- *La propension de l'IA à « halluciner », c'est-à-dire à générer des informations fausses ou trompeuses, accroît la diffusion d'informations erronées, compromettant ainsi davantage la fiabilité et la crédibilité des contenus en ligne.*
- *La réduction des coûts et des compétences requises pour la production de désinformation, y compris les deepfakes, abaisse les barrières pour les acteurs malveillants.*
- *Les biais des algorithmes d'IA peuvent fausser la création de contenu, la diffusion et la prise de décision, et donc perpétuer les préjugés et renforcer les stéréotypes et les inégalités.*
- *Les risques accrus de violation des droits d'auteur, de la vie privée et des lois contre la diffamation mettent en péril la propriété intellectuelle, les données personnelles et les réputations.*

D'un autre côté, l'IA offre des possibilités réelles de démocratisation de l'espace informationnel et d'amélioration de la fiabilité de l'écosystème de l'information. Nombre de ces possibilités, qui ne répondent généralement pas aux priorités de maximisation des revenus, de valeur actionnariale et de monopolisation du marché, restent largement inexploitées.

Les équipes de conception et d'ingénierie de la poignée d'entreprises et d'entités spécialisées dans l'IA qui conçoivent et améliorent les systèmes disposent donc d'une occasion unique de définir l'avenir de l'IA au niveau le plus fondamental, au travers des phases de conception, de développement et de déploiement. En optant dès le départ pour des approches responsables et en veillant à ce que leurs équipes soient le reflet de la diversité du monde dans lequel leurs produits fonctionneront, ils sont en mesure de réduire considérablement les risques et de prévenir les incidences négatives dans l'espace de l'information et de la communication. À cette fin, les valeurs humaines, l'éthique et la diversité doivent être au cœur du développement et du déploiement de l'IA, et y être intégrés tout au long du cycle de vie de l'IA. Des mesures proactives d'atténuation des effets négatifs potentiels doivent également être prises. Tout ceci implique d'aborder l'IA comme un bien public et non comme une simple innovation destinée à générer des profits pour les actionnaires.

Il est primordial d'adopter une approche exhaustive pour l'évaluation des systèmes d'IA. Cette démarche implique des exercices en red-teaming afin d'identifier et de traiter les vulnérabilités aux dommages potentiels, y compris, mais pas uniquement, la mésinformation, la manipulation et l'utilisation malveillante. Seules de minutieuses évaluations des risques peuvent permettre d'évaluer les biais de l'IA et ses conséquences sur la vie privée, en veillant au respect des normes éthiques et juridiques internationales, y compris les lois internationales sur les droits humains et le droit international humanitaire[66]. Enfin, un suivi après le déploiement est indispensable pour adapter les systèmes d'IA à l'évolution du paysage informationnel, pour garantir des représentations mondiales de la diversité et pour constamment mettre à jour les systèmes d'IA afin de prévenir les risques émergents.

66 À partir de ce point, le terme « droit international des droits humains » désignera à la fois le droit international des droits humains et le droit international humanitaire.

1. MISE EN PLACE DE SYSTÈMES D'IA POUR LA CRÉATION, LA DISTRIBUTION ET LA CONSOMMATION D'INFORMATIONS EN TOUTE SÉCURITÉ ET DANS UNE OPTIQUE INCLUSIVE

Les développeurs et les déployeurs de systèmes d'IA exercent une influence considérable sur l'espace de l'information et de la communication. Ils ont de lourdes responsabilités, notamment en ce qui concerne les conséquences sur les droits humains fondamentaux tels que le droit à la vie privée, la liberté d'opinion et d'expression, l'égalité et la non-discrimination, l'accès à l'information et la liberté de la presse. La résolution de l'Assemblée générale des Nations unies sur la *Promotion et la protection des droits humains dans le contexte des technologies numériques*, adoptée en 2023, stipule « qu'il faut respecter, protéger et promouvoir les droits humains et les libertés fondamentales tout au long du cycle de vie des systèmes d'intelligence artificielle et que les technologies numériques nouvelles et émergentes devraient fournir de nouveaux moyens de promouvoir, de protéger et d'exercer les droits humains et non d'y porter atteinte »[67].

Les équipes de conception et d'ingénierie doivent être la première ligne de défense des droits humains et du droit à une information fiable, tel que défini par le Partenariat pour l'information et la démocratie[68]. Les concepteurs et les ingénieurs doivent développer des systèmes d'IA au service des valeurs démocratiques et de l'intérêt public. Cela suppose une vision nuancée des risques que les systèmes d'IA représentent pour un espace mondial de l'information pluriel, libre et digne de confiance, ainsi qu'un engagement en faveur d'un apprentissage et d'une adaptation continus.

Dans ce contexte complexe, la mise en œuvre de solutions techniques pour répondre au risque de partialité doit être une priorité absolue. Les entreprises forment les systèmes d'IA à partir de données historiques, ce qui risque de perpétuer les préjugés et de violer le droit à la non-discrimination fondée sur la race, la couleur, le sexe, l'orientation sexuelle, l'identité de genre, la langue, la religion, l'opinion politique ou autre, l'origine nationale ou sociale, la propriété, la naissance ou tout autre statut[69].

Il est tout aussi crucial de lutter contre l'hégémonie culturelle en matière de données d'apprentissage. Les systèmes d'IA ne devraient pas seulement refléter la vision des grands centres technologiques, mais aussi les points de vue et les langues de diverses cultures et communautés de « la majorité globale ». Pour ce faire, les expériences, valeurs et actions de la communauté mondiale au sens large doivent impérativement être intégrées dans l'apprentissage des systèmes d'IA. L'inclusivité dans l'IA ne garantit pas seulement sa pertinence à l'échelle mondiale, mais favorise également le multilinguisme et l'égalité

67 Assemblée Générale des Nations Unis (2023). Promotion et protection des droits humains dans le contexte des technologies numériques. Disponible sur : https://digitallibrary.un.org/record/4032837?ln=fr&v=pdf (Consulté le 7 février 2024).
68 Forum sur l'information et la démocratie. Partenariat International sur l'Information et la Démocratie. Disponible sur : https://informationdemocracy.org/fr/partenariat-international-information-democratie/ (Consulté le 8 février 2024).
69 Organisation des Nations Unies (2001). Déclaration universelle des droits de l'homme, Article 2. Disponible sur : https://www.un.org/fr/about-us/universal-declaration-of-human-rights (Consulté le 7 février 2024).

d'accès à la connaissance. Cette approche est essentielle à la préservation de la diversité culturelle et à l'équité entre les différentes cultures pour qu'elles soient reflétées dans les systèmes d'IA[70].

Par ailleurs, il est également essentiel de prendre en compte la protection de la vie privée et de la propriété intellectuelle, et ce dès la phase de conception et tout au long du cycle de vie de l'IA. Les données étant une ressource primordiale, il est fondamental de maintenir la transparence de leur utilisation, de garantir une sécurité rigoureuse et de respecter les lois relatives à la propriété intellectuelle et à la protection des données.

Les développeurs doivent veiller en priorité à protéger l'IA contre les utilisations abusives et les attaques les plus complexes. Il leur appartient de prendre des mesures les plus rigoureuses pour vérifier la provenance et l'authenticité des données, condition *sine qua non* à l'amélioration de la transparence des systèmes d'IA et à leur résistance à la désinformation. En outre, en ce qui concerne les menaces sophistiquées, les développeurs doivent mettre en place des stratégies d'atténuation. Ces dernières devraient inclure des mesures visant à accroître la difficulté et le coût de toute tentative de manipulation de l'IA susceptible de porter atteinte aux droits humains.

Divers acteurs internationaux ont élaboré des principes éthiques pour guider les développeurs et les déployeurs d'IA dans la conception, le développement et la mise en œuvre de ces systèmes. Parmi ces principes figurent, entre autres, le *Décret sur le développement et l'utilisation sûrs, sécurisés et dignes de confiance de l'intelligence artificielle* signé par le président des États-Unis Biden, la loi de l'UE sur l'IA (qui devrait être finalisée au printemps 2024), la *Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle*[71], qui comprend un chapitre politique expressément consacré à la communication et à l'information, fournissant des recommandations politiques concrètes et spécifiques au domaine, les *Principes d'Asilomar sur l'IA*[72], la *Déclaration de Windhoek sur l'intelligence artificielle en Afrique australe*[73], le *Projet de Convention-cadre sur l'Intelligence Artificielle Droits de l'homme, démocratie et État de droit* du Conseil de l'Europe[74], la *Recommandation du Conseil de l'OCDE sur l'intelligence artificielle*[75], les *Principes directeurs internationaux du processus Hiroshima pour les systèmes d'IA avancés*[76], et le *Guide de l'ASEAN sur l'éthique et la gouvernance de l'IA*[77]*; et les Lignes directrices en matière d'éthique pour une IA digne de confiance élaborées par le groupe d'experts de haut niveau de l'UE sur l'IA*[78]*.* Des principes éthiques spécifiques à chaque pays pourraient également être définis pour traduire les valeurs sociales et juridiques, sous réserve qu'ils soient conformes à la législation, aux normes et aux règles internationales en matière de droits humains.

70 Organisation des Nations Unies. *Déclaration Universelle sur la Diversité Culturelle,* Article 6. Disponible sur : www.ohchr.org/en/instruments-mechanisms/instruments/universal-declaration-cultural-diversity#:~:text=Freedom%20of%20expression%2C%20media%20pluralism,the%20guarantees%20of%20cultural%20diversity (Consulté le 7 février 2024).

71 UNESCO (2021). Recommandation sur l'éthique de l'intelligence artificielle. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Consulté le 7 février 2024).

72 Future of Life (2017). *Asilomar AI Principles*,. Disponible sur : https://futureoflife.org/open-letter/ai-principles/ (Consulté le 7 février 2024).

73 UNESCO (2022). Déclaration de Windhoek sur l'intelligence artificielle en Afrique australe Windhoek. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000383197_fre (Consulté le 7 février 2024).

74 Conseil européen (2024). Législation sur l'intelligence artificielle: le Conseil et le Parlement parviennent à un accord sur les premières règles au monde en matière d'IA. Disponible sur : https://www.consilium.europa.eu/fr/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ (Consulté le 7 février 2024).

75 OCDE (2019). Recommandation du Conseil sur l'intelligence artificielle. Disponible sur : https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449 (Consulté le 7 février 2024).

76 Commission Européenne (2023).Code de conduite international pour les systèmes d'IA avancés dans le cadre du processus Hiroshima. Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Consulté le 7 février 2024).

77 ASEAN (2023). ASEAN Guide on AI Governance and Ethics. Disponible sur : https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf (Consulté le 7 février 2024).

78 Commission Européenne (2019). Lignes directrices en matière d'éthique pour une IA digne de confiance. Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/ethics-guidelines-trustworthy-ai (Consulté le 15 février 2024).

RECOMMANDATIONS AUX **ÉTATS**

> **Proposer une assistance pratique aux développeurs et déployeurs d'IA pour transformer le droit international et les principes éthiques de l'IA[79] en mesures concrètes de promotion et de protection des droits humains.** Cela suppose de favoriser le respect des règles en fournissant des lignes directrices détaillées et des exemples concrets, qui expliquent comment mettre en pratique avec succès les principes existants en matière de conception éthique.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA[80]**

> **Prendre en compte les considérations relatives à l'impact sociétal au sens large dans les processus de prise de décision concernant la conception, le développement et le déploiement de systèmes d'IA dans l'espace de l'information et de la communication.** Cette approche, qui pourrait être facilitée par des outils tels que l'Évaluation de l'impact éthique de l'UNESCO[81], est essentielle pour contrebalancer l'importance accordée à la croissance, à la maximisation des profits et à la recherche d'avantages concurrentiels.

1.1 JEUX DE DONNÉES D'APPRENTISSAGE

L'intégrité et la qualité de tout système d'IA dépendent essentiellement des données à partir desquelles il apprend, ainsi que des personnes qui identifient et fournissent ces données. L'introduction potentielle de biais reste très préoccupante au cours du processus de développement d'un système d'IA, en particulier lors de la collecte et du prétraitement des données. À ce stade, les biais peuvent apparaître sous diverses formes, notamment le biais d'étiquetage, le biais d'échantillonnage et le biais d'exclusion qui résulte de la suppression ou de l'omission inopportune de données pertinentes dans le jeu de données[82]. De tels biais dans la définition et la modélisation du jeu de données d'apprentissage peuvent avoir de sérieuses répercussions. Les biais sont encore accentués lorsque les équipes de développeurs qui forment ces systèmes manquent de diversité, notamment en termes cognitifs, religieux, ethniques, de genre, de sexe, d'origine géographique et économiques.

En premier lieu, les biais introduits dans les systèmes d'IA par l'incorporation de données d'apprentissage non représentatives peuvent entraîner une marginalisation accrue des communautés marginalisées. Par exemple, dans le cadre de la modération de contenu sur les plateformes de réseaux sociaux, si les données d'apprentissage sont principalement composées de contenus signalés comme inappropriés par des utilisateurs issus de milieux culturels spécifiques, l'IA peut développer une compréhension biaisée

79 Parmi les principes éthiques de l'intelligence artificielle reconnus à l'échelle mondiale on retrouve ceux inclus dans la Recommandation sur l'Éthique de l'Intelligence Artificielle de l'UNESCO , ainsi que d'autres principes éthiques énumérés dans l'introduction.

80 Tout au long du rapport, la terminologie « entreprises et entités d'IA » est utilisée en référence aux acteurs qui développent et déploient les systèmes d'IA, et ce même si ce n'est pas leur activité principale.

81 UNESCO (2023). *Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence.* Disponible sur : www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence (Consulté le 7 février 2024).

82 Parlement Européen (2022). *Auditing the quality of datasets used in algorithmic decision-making systems.* Disponible sur : www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf (Consulté le 7 février 2024).

de ce qui constitue un contenu offensant. Cette interprétation biaisée pourrait conduire à une censure excessive de sujets ou d'expressions pertinents pour les groupes sous-représentés, sans pour autant identifier et traiter les contenus préjudiciables à ces communautés. Pour remédier à ce phénomène, il est essentiel d'inclure dans les jeux de données d'apprentissage tout un éventail d'expressions et de nuances culturelles au sein d'une même langue. Pour ce faire, il convient au minimum d'assurer la diversité des points de vue, des expériences et des origines des membres des équipes concernées.

En outre, les biais des systèmes d'IA peuvent résulter des limites inhérentes aux outils de traitement du langage naturel, souvent développés pour une seule langue et susceptibles de ne pas fonctionner aussi efficacement lorsqu'ils sont utilisés dans d'autres dialectes ou langues. Cette problématique est particulièrement épineuse pour les langues dont l'empreinte en ligne est plus réduite et dont les données d'apprentissage disponibles sont limitées ou ne reflètent que partiellement la société ou la culture. Dans la mesure où les algorithmes ne disposent pas de suffisamment de données pour ces langues, leur capacité d'apprentissage est réduite, ce qui se traduit par des outils automatisés présentant des taux d'erreur plus élevés[83]. Malheureusement, cette situation renforce l'hégémonie culturelle des régions et des langues dominantes. Il est impératif de remédier à ce manque de données afin de combler le fossé numérique entre les différentes régions du monde, de développer des systèmes d'IA adaptés aux besoins des pays de « la majorité globale » et des communautés linguistiques plus réduites, et de redistribuer le pouvoir technologique de manière plus équitable. Si les modèles linguistiques multilingues sont prometteurs, ils comportent aussi d'importantes limites. Parmi celles-ci, citons le recours à des textes traduits par des machines, souvent émaillés d'erreurs, et l'absence de références contextuelles[84]. La mise en œuvre prématurée de ces modèles et l'absence de garanties suffisantes peuvent compromettre la liberté d'expression au sein de certaines régions ou de certains groupes en raison de l'inexactitude du filtrage et du marquage d'informations spécifiques, comme c'est le cas pour les systèmes de modération de contenu. Certaines démocraties non anglophones sont confrontées à un risque plus élevé en raison de la disponibilité limitée des données, de même que certains pays où la liberté d'expression reste menacée ou est sérieusement entravée par des régimes autoritaires ou peu démocratiques[85]. Il convient de souligner que ce risque augmente considérablement dans les pays de « la majorité globale », et plus encore dans les régions où sont parlés des langues et des dialectes moins répandus.

Par ailleurs, plus les modèles d'IA prennent de place dans différents aspects de notre vie, plus le défi que représente la protection de la vie privée est important. L'adoption grandissante de l'IA et son évolution permanente nécessitent une collecte de volumes de données plus importants (cet aspect majeur sera examiné plus en détail dans les sous-sections a. et e.), ce qui peut engendrer des méthodes de collecte de données plus intrusives, y compris des techniques de surveillance renforcées et des conceptions d'interface utilisateur manipulatrices. De telles pratiques exacerbent les préoccupations relatives à l'utilisation et à la sécurité des informations personnelles au sein des systèmes d'IA, représentant un risque important pour la vie privée des individus. Ainsi, les systèmes d'IA générative étant formés sur des jeux de données comprenant des données personnelles sensibles, notamment des identifiants personnels, l'orientation sexuelle ou les affiliations politiques (entre autres), ils sont susceptibles de produire des résultats exposant involontairement ces informations confidentielles.

Ce problème est aggravé par le fait que les avancées technologiques dans le domaine de l'IA sont si rapides qu'elles remettent en question, voire dépassent, les cadres juridiques actuels en matière de protection des données, lorsqu'ils existent.

---

83 Díaz, A. and Hecht-Felella, L. (2021). *Double Standards in Social Media Content Moderation, Brennan Center for Justice*. Disponible sur : www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf (Consulté le 7 février 2024).

84 Nicholas, G. and Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis,* Center for Democracy and Technology. Disponible sur : https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Consulté le 7 février 2024).

85 *Ibid*

Enfin, il est difficile de savoir si les techniques actuelles de collecte de données sont conformes aux approches juridiques en vigueur en matière de propriété intellectuelle. La non-conformité en la matière présente des risques non seulement pour les droits individuels, mais aussi pour la soutenabilité du journalisme.

A. CURATION DE DONNÉES INCLUSIVE ET TRANSPARENTE

Les équipes doivent faire des choix lorsqu'elles forment des jeux de données. Ces choix en matière de provenance et de curation des jeux de données conditionnent la qualité et les résultats des systèmes d'IA. Compte tenu de l'influence considérable que ces résultats exercent sur l'espace informationnel, et du risque d'atteintes aux droits humains qu'ils représentent, ces décisions ne devraient pas être du seul ressort des ingénieurs et des entreprises et entités spécialisées dans l'IA. De plus, elles ne devraient pas non plus être confiées à des équipes caractérisées par un manque de diversité interne. Il est urgent que la société civile, les chercheurs et les communautés concernées s'impliquent de manière durable et significative. Une approche inclusive est indispensable pour garantir que les choix relatifs aux jeux de données reflètent des valeurs sociétales plus diverses, et pas uniquement des priorités techniques ou d'entreprise, ni le parti pris involontaire de groupes d'ingénieurs homogènes. Cette démarche ne doit pas inciter les entreprises et les entités spécialisées dans l'IA à se soustraire à leur responsabilité de respecter les principes éthiques relatifs à la provenance et à la curation des données, mais doit au contraire permettre aux différents partenaires de constituer des jeux de données d'apprentissage.

En outre, les jeux de données d'apprentissage doivent pouvoir faire l'objet de contrôles externes[86] afin de vérifier leur conformité avec les lois et règlements en vigueur, d'identifier les problèmes potentiels et de permettre aux ayants droit cherchant un recours en cas de suspicion d'utilisation abusive de leurs données de rassembler des preuves et d'intenter une action en justice. Ainsi, les entreprises et les entités spécialisées dans l'IA devraient garantir une transparence à plusieurs niveaux sur les données qu'elles utilisent. Si les chercheurs indépendants, les organismes de contrôle et les autres entités certifiées doivent avoir un accès direct aux jeux de données d'apprentissage à des fins d'investigation (voir chapitre 4, section 4.2), le grand public doit être en mesure de comprendre certaines caractéristiques fondamentales des jeux de données d'apprentissage dans un format simple et accessible.

RECOMMANDATIONS
AUX **ÉTATS**

> **Imposer aux entreprises et aux entités spécialisées dans l'IA de rendre publiques les informations relatives à leurs jeux de données d'apprentissage, de façon accessible et compréhensible, y compris sous la forme d'une base de données consultable.** Ces informations doivent comprendre :
>
> - La provenance et composition des données, y compris la manière dont celles-ci ont été obtenues et sélectionnées, et si elles font l'objet d'une licence. Ces informations portent également sur l'utilisation de données protégées par des droits d'auteur ou d'autres droits de protection légaux.
> - Un récapitulatif des caractéristiques démographiques des personnes dont les données sont incluses dans le jeu de données.

86 Selon la proposition de règlement de l'UE sur l'IA, les autorités nationales compétentes devraient avoir pleinement accès aux jeux de données d'entraînement, de validation et de test utilisés pour le développement des systèmes d'IA à haut risque (article 63.7a). En outre, dans des cas particuliers, le code source du système d'IA à haut risque peut également être demandé (article 63,7 b).

- Les métadonnées sur les sujets, les contextes culturels et les langues couverts par le jeu de données, y compris les informations sur le volume et le format des données.
- Les limites et les biais potentiels connus inhérents aux jeux de données.
- Une description des étapes du traitement des données, y compris la manière dont les données ont été nettoyées et préparées pour l'apprentissage, ainsi que les techniques d'anonymisation et d'agrégation utilisées.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Mettre en place un processus inclusif et participatif, comprenant à minima une participation équitable, durable et substantielle de chercheurs indépendants, de la société civile et des communautés concernées. Ce processus devrait être utilisé pour déterminer les règles et les critères régissant la provenance et la curation des jeux de données[87].** Cela implique en particulier d'établir des critères d'inclusion ou d'exclusion de données potentiellement nuisibles et trompeuses dans les jeux de données d'apprentissage pour les modèles de base. Par ailleurs, il est nécessaire de veiller à ce que ces processus, règles et critères soient transparents pour le grand public.

> **Respecter les principes éthiques et les lois en matière de provenance des données et de pratiques de curation, qui exigent au minimum :**

- **Le respect de la gouvernance des données, de la protection des données et des lois sur la propriété intellectuelle pour les choix de provenance et de curation.**
- **La garantie que les ensembles de données traités ne fassent pas l'objet d'une discrimination** fondée sur la race, la couleur, le sexe, l'orientation sexuelle, l'identité de genre, la langue, la religion, l'opinion politique ou autre, l'origine nationale ou sociale, la propriété, la naissance ou tout autre statut.

RECOMMANDATIONS AUX **ÉTATS ET AUX ORGANISMES INTERNATIONAUX ET MULTILATÉRAUX DE NORMALISATION**

> **Organiser des discussions multipartites afin d'établir des normes et des lignes directrices concernant les types de données autorisées pour l'apprentissage des modèles d'IA,** y compris les sujets sensibles tels que les données relatives aux enfants, et les catégories potentiellement sensibles telles que l'historique des recherches et des déplacements qui incluent des éléments tels que les visites dans des cliniques de santé spécialisées, les endroits fréquentés par la communauté LGBTQ+, les bureaux des organismes de défense des droits humains ou des partis politiques, et bien plus encore. Ces discussions devraient également porter sur les politiques publiques en matière d'apprentissage par renforcement et de red-teaming.

87 Jernite, Y; (2023). *Training Data Transparency in AI: Tools, Trends, and Policy RECOMMANDATIONS*, Hugging Face Community Blog. Disponible sur : https://huggingface.co/blog/yjernite/data-transparency#data-transparency-in-focus-what-is-needed (Consulté le 7 février 2024).

B. ACCÈS À DES JEUX DE DONNÉES DE HAUTE QUALITÉ

La disponibilité de données de haute qualité est un impératif stratégique qui conditionne le développement responsable de l'IA et constitue le fondement d'une innovation inclusive.

Les États ont un rôle central à jouer dans ce domaine, en encourageant le développement de jeux de données de haute qualité dans le secteur public, en les publiant selon des normes de données ouvertes et en soutenant le développement de jeux de données au sein des régions, des cultures et des langues sous-représentées.

Les pratiques actuelles de curation des données mettent également en évidence la nécessité de mettre en place un cadre solide de gouvernance des données, apportant des directives complètes sur l'extraction, la collecte et l'utilisation des données. Ce cadre doit privilégier l'intérêt public et traiter les problématiques complexes liées aux droits d'auteur et à la protection de la vie privée. En outre, les autorités publiques devraient aider les propriétaires et les utilisateurs de données à rationaliser les procédures de partage de ces dernières, en mettant à leur disposition des outils et des mécanismes normalisés qui garantissent la loyauté et l'équité des échanges de données.

RECOMMANDATIONS AUX **ÉTATS**

> **Favoriser le développement et la disponibilité de jeux de données du secteur public de haute qualité et la mise en œuvre de normes de données ouvertes.** Cela inclut la production de données, le nettoyage des jeux de données et leur mise à disposition conformément aux normes de données ouvertes dans un format lisible et réutilisable par machine[88].

> **Encourager le développement de jeux de données au sein de pays, de cultures et de langues sous-représentés en affectant des ressources et en fournissant un soutien. Cela comprend les subventions de recherche, le financement de la société civile, les appels à contribution, les dons de données[89] et le soutien aux médias en tant que créateurs de données (voir la sous-section f).** La priorité doit être donnée à l'inclusion d'un large éventail de voix et à la représentation des communautés historiquement marginalisées, en veillant à ce que les systèmes d'IA reflètent la diversité des cultures et des valeurs présentes dans le pays ou la communauté.

> **Définir un cadre prévoyant des dispositions explicites en matière de gouvernance des données. Celui-ci devrait fixer des règles claires concernant l'extraction, la collecte et l'utilisation des données.** L'utilisation des données et l'accès aux jeux de données devraient se fonder sur l'intérêt que présente le projet d'IA pour le public, sur la qualité, la légalité et la sensibilité des jeux de données, ainsi que sur la quantité de données concernée. La présomption d'illégalité devrait s'appliquer à toute extraction de données à des fins d'inclusion dans des jeux de données d'apprentissage, à moins que les données ne soient publiées sous des

88 *Les principes FAIR (FAIR Guiding Principles) pour la gestion et la conservation des données scientifiques peuvent fournir un cadre pour la publication de telles ensembles de données.* Disponible sur : www.nature.com/articles/sdata201618 (Consulté le 7 février 2024).

89 Les « dons de données » font référence à une pratique dans laquelle des individus, des organisations publiques, des entreprises ou des entités fournissent volontairement leurs données à des fins de recherche, de développement et d'autres utilisations dans l'intérêt public. Dans le contexte du développement de l'IA, ces dons de données peuvent aider à enrichir des ensembles de données, en particulier pour les pays, les cultures, les identités et les langues sous-représentés.

normes de données ouvertes ou que les développeurs et déployeurs d'IA ne puissent justifier de la légalité de leur utilisation de jeux de données. Cette démarche nécessite également une mise à jour des définitions des licences de partage de données afin d'assurer la transparence juridique, telles que les licences de type Creative Commons (élaborées avant la généralisation des pratiques d'extraction de données par l'IA). En règle générale, la minimisation des données devrait gouverner les activités des entreprises technologiques, en ce sens qu'elles ne devraient collecter que les données minimales nécessaires au fonctionnement de leurs produits.

> **Élaborer des mécanismes de partage des données pour les structures détentrices de données et celles qui cherchent à les utiliser pour développer l'IA[90].** Afin que les données puissent être partagées, leurs propriétaires et leurs consommateurs sont souvent amenés à négocier les conditions au cas par cas. L'élaboration d'outils, de modèles et de mécanismes pratiques normalisés de partage des données permettrait non seulement de surmonter les obstacles pratiques, mais également de clarifier la procédure d'attribution de propriété ou de droits liés aux résultats générés à partir de ces données partagées (p. ex. modèles d'IA, perspectives analytiques, résultats de recherche) et de répartir les bénéfices. En définitive, ces cadres garantiraient des échanges de données sûrs, équitables et justes pour toutes les parties concernées. La confiance étant cruciale dans les pratiques de partage des données, ces cadres devraient être définis et gérés par un organisme indépendant jouissant d'une réputation bien établie en matière de sécurité et d'efficacité des données. Celui-ci jouerait un rôle central en exerçant des fonctions clés telles que la mise en place d'un système de définition des données à partager, la médiation des objectifs et de l'utilisation prévue des données, la création de protocoles de transfert et de stockage des données ainsi que la clarification de la répartition de la valeur commerciale générée.

> **Faciliter l'accès au matériel protégé par des droits d'auteur afin de l'utiliser comme données d'apprentissage pour la recherche d'intérêt public, tout en protégeant la propriété intellectuelle et la vie privée des utilisateurs.**

> **Encourager les entreprises et les entités spécialisées dans l'IA à rendre leurs données d'apprentissage accessibles à la recherche d'intérêt public et aux alternatives aux systèmes d'IA à but lucratif.**

C. LA QUESTION DES BIAIS

L'apprentissage sur lequel s'appuient les systèmes d'IA est souvent formé de vastes jeux de données reflétant principalement des approches anglocentriques et eurocentriques. En outre, les communautés historiquement marginalisées sont souvent sous-représentées ou mal représentées au sein de ces jeux de données, mais aussi au sein des équipes qui collectent les données et forment les systèmes. Le manque de diversité de ces données et de ces équipes engendre des biais qui se répercutent sur les systèmes d'IA.

Ces partis pris systématiques peuvent se traduire par un manque de diversité des perspectives et des contenus que ces systèmes comprennent et génèrent. Dans certains cas, ces biais peuvent aussi produire des résultats préjudiciables qui affectent de manière disproportionnée les cultures sous-représentées,

90 Cette recommandation se base sur la proposition de Data Trusts présentée au gouvernement britannique : Wendy Hall, D. and Pesenti, J. (2017). Growing the Artificial Intelligence Industry in the UK. Disponible sur : https://assets.publishing.service.gov.uk/media/5a824465e5274a2e87dc2079/Growing_the_artificial_intelligence_industry_in_the_UK.pdf (Consulté le 7 février 2024).

les systèmes ne discernant pas correctement la diversité des contextes et nuances culturels ou n'y réagissant pas de manière appropriée. Par exemple, il a été constaté que la surreprésentation des points de vue occidentaux dans les jeux de données peut conduire les systèmes d'IA à perpétuer des stéréotypes et à générer des contenus inappropriés ou offensants, en particulier dans la représentation des femmes et des cultures non occidentales, les résultats reflétant souvent une hyper sexualisation, ou des idéologies misogynes, voire suprémacistes blanches[91].

Dans ce contexte, il est impératif de remédier aux biais des données d'apprentissage pour contrer le risque d'hégémonie de la culture dominante.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Veiller à ce que les équipes travaillant sur les ensembles de données d'apprentissage reflètent l'inclusion et la diversité.** Cela suppose de promouvoir une main-d'œuvre diversifiée, de former le personnel aux préjugés systémiques, de renforcer la diversité des filières de recrutement et de collaborer avec les instances éducatives afin d'encourager des personnes de tous horizons à rejoindre le secteur.

> **Enrichir les jeux de données d'apprentissage pour aborder les questions de biais et de représentation des différentes cultures. Cela inclut :**

- **La diversification des données utilisées pour la formation des systèmes d'IA en collectant des données authentiques émanant de groupes, de régions, de cultures et de langues sous-représentés.** Cette approche s'appuie sur l'obtention directe de données auprès de groupes, régions, cultures et langues actuellement sous-représentés dans les jeux de données, moyennant une compensation équitable et des pratiques éthiques. L'une des stratégies envisageables pour faciliter cette collecte consiste à organiser des « campagnes de dons de données » offrant aux communautés la possibilité de contribuer à des projets d'intérêt public, tout en leur garantissant une compensation adéquate pour leurs précieuses données.
- **La pondération différenciée des segments de données.** Pondérer les différents segments de données de manière appropriée permet d'équilibrer le jeu de données durant l'apprentissage, en accordant plus d'importance aux segments sous-représentés ou plus sensibles. Une compréhension approfondie des implications de cet ajustement des paramètres est essentielle afin de garantir la réduction des biais sans affecter négativement d'autres critères de performance.
- **La génération de données synthétiques.** Grâce à des techniques artificielles, telles que les réseaux antagonistes génératifs (GAN) et les techniques d'augmentation des données[92], il est possible de créer un jeu de données d'apprentissage plus représentatif et plus proche de la distribution réelle des données, notamment pour les classes minoritaires dans les jeux de données déséquilibrés. Les données synthétiques sont particulièrement utiles dans les cas où les données existantes sont biaisées ou manquent de diversité. Bien que cette méthode permette d'atténuer les problèmes de confidentialité, de droits d'auteur et d'éthique liés à l'utilisation de données réelles, elle doit cependant être utilisée

91 Birhane, A. et al. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. Disponible sur : https://arxiv.org/abs/2110.01963 (Consulté le 7 février 2024).

92 Shorten, C. and Khoshgoftaar, T.M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1). Disponible sur : https://doi.org/10.1186/s40537-019-0197-0. (Consulté le 14 février 2024).

avec précaution, car si les données synthétiques peuvent en effet contribuer à corriger les biais et à renforcer la protection de la vie privée[93], une dépendance excessive à ce type de données pourrait également nuire à la fiabilité des systèmes d'IA.[94]

- **L'audit et le calibrage des jeux de données.** Cette étape consiste à comprendre la composition des jeux de données et à les ajuster de façon à ce qu'ils reflètent davantage de contextes culturels et démographiques. Ce processus permet de minimiser les biais susceptibles d'être introduits dans l'algorithme.[95][96]
- **Des investissements proactifs dans le développement de solutions techniques pour constituer des jeux de données reflétant fidèlement la diversité des cultures du monde entier,** mettant l'accent sur l'inclusion des groupes sous-représentés et des populations plus restreintes, et ce afin de contrecarrer les biais inhérents aux jeux de données de référence standard.

> **Procéder à des évaluations d'impact pour vérifier l'absence de biais, y compris en matière de diversité et de représentation, avant le déploiement des systèmes d'IA, et les réexaminer en permanence.** Pour cela, il est possible de faire appel à des « red teams » internes ou à des tiers chargés d'auditer les données et les modèles[97].

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux entreprises et aux entités spécialisées dans l'IA de procéder à des analyses d'impact afin de vérifier l'existence de biais, y compris en matière de diversité et de représentation, avant le déploiement des systèmes d'IA, et de les réexaminer en permanence, ce qui implique également de vérifier la présence de biais dans le cadre d'audits indépendants** (voir chapitre 4, section 2.3) **et de permettre un examen minutieux par les chercheurs** (voir chapitre 4, section 4.2).

D. LANGUES À FAIBLE NIVEAU DE RESSOURCES

Ces derniers temps, les entreprises spécialisées dans l'IA rivalisent pour intégrer le plus grand nombre de langues dans leurs modèles linguistiques multilingues. Cependant, une question essentielle reste sans réponse : comment les modèles linguistiques peuvent-ils gérer la « malédiction du multilinguisme », c'est-à-dire le fait que pour une taille de modèle fixe, l'ajout de données multilingues entraîne une dégradation des performances du modèle linguistique, tant pour les langues à faibles ressources que pour les langues à ressources élevées[98]. Ainsi, les entreprises qui développent de grands modèles linguistiques doivent se concentrer non seulement sur le nombre de langues incluses, mais également sur la qualité des performances du modèle dans chaque langue[99].

93 Zewe, A. (2022). In machine learning, synthetic data can offer real performance improvements, MIT News. Disponible sur : https://news.mit.edu/2022/synthetic-data-ai-improvements-1103 (Consulté le 7 février 2024).
94 Marwala, T. (2023). Algorithm Bias — Synthetic Data Should Be Option of Last Resort When Training AI Systems, United Nations University. Disponible sur https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems (Consulté le 7 février 2024).
95 Cambridge Consultants (2019). Use of AI in Online Content Moderation. Disponible sur : www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Consulté le 7 février 2024).
96 European Parliament (2022). Auditing the quality of datasets used in algorithmic decision-making systems. Disponible sur : www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf (Consulté le 7 février 2024)
97 Silberg, J. and Manyika J. (2019). Notes from the AI frontier: Tackling bias in AI (and humans), McKinsey Global Institute. Disponible sur : www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans (Consulté le 7 février 2024).
98 Chang, T.A., Arnett, C., Tu, Z. and Bergen, B.K., (2023). When is multilinguality a curse? Language modeling for 250 high-and low-resource languages, arXiv. Disponible sur : https://doi.org/10.48550/arXiv.2311.09205. (Consulté le 7 février 2024).
99 Nicholas, G. and Bhatia, A. (2023). Lost in Translation: Large Language Models in Non-English Content Analysis, Center for Democracy and Technology. Disponible sur : https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Consulté le 7 février 2024).

Le développement des capacités de traitement du langage naturel (TALN) dans les langues à ressources élevées comme l'anglais constitue un cercle vertueux : des jeux de données abondants, propres et annotés par l'homme permettent d'élaborer des modèles et des repères plus avancés, ce qui favorise la recherche, les publications et les applications pratiques, qui à leur tour stimulent la demande de jeux de données encore plus nombreux. À l'inverse, les langues à faible niveau de ressources sont prisonnières d'un cercle vicieux en raison du manque d'outils, d'annotateurs, de financement et de reconnaissance dans les publications et les communautés de TALN dominantes. Afin d'initier un cercle vertueux pour les langues à faible niveau de ressources, les investissements devraient se concentrer sur la promotion de communautés scientifiques de TALN autonomes, ce qui suppose des efforts simultanés à tous les niveaux[100].

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Procéder à des analyses d'impact afin d'identifier d'éventuelles inexactitudes et représentations erronées dans les différentes langues dans lesquelles le système d'IA sera disponible, avant son déploiement, et les réexaminer en permanence.** Pour cela, il est possible de faire appel à des « red teams » internes ou à des tiers chargés d'auditer les données et les modèles.

> **Définir des mesures et des critères de référence clairs pour évaluer les performances des modèles linguistiques dans chaque langue.**

> **Mettre au point des techniques de création de modèles linguistiques plus performants sur le plan des données, et dont le fonctionnement nécessiterait moins de ressources, y compris :**

- En concevant des modèles mettant l'accent sur des domaines ou des tâches spécifiques[101].
- En exploitant les modèles multilingues pour l'apprentissage par transfert des langues à ressources élevées vers les langues à faible niveau de ressources.
- En définissant des critères de performance multilingues.

> **Favoriser le développement de communautés scientifiques autonomes dans le domaine du TALN par le biais de subventions et le partage des jeux de données linguistiques à faibles ressources utilisés pour l'apprentissage des modèles d'IA, y compris (au moins en partie) des données d'apprentissage exclusives[102].**

100 *Ibid*
101 Miller, K and Lohn, A. (2023). Techniques to Make Large Language Models Smaller: AN Explainers, Center for Security and Emerging Technologies. Disponible sur : https://cset.georgetown.edu/publication/techniques-to-make-large-language-models-smaller-an-explainer/ (Consulté le 7 février 2024).
102 Nicholas, G. and Bhatia, A. (2023). Lost in Translation: Large Language Models in Non-English Content Analysis, Center for Democracy and Technology. Disponible sur : https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Consulté le 7 février 2024).

RECOMMANDATIONS
AUX **ÉTATS**

> **Investir dans la recherche sur les modèles d'IA pour les langues à faible niveau de ressources afin de remédier aux lacunes actuelles du marché.** Les incitations gouvernementales peuvent encourager les chercheurs en TALN à diversifier leur champ d'action entre différentes langues et approches, plutôt que de se focaliser principalement sur l'anglais, et à mesurer l'impact des modèles d'IA sur les différentes communautés linguistiques, la préservation des langues et le renforcement de la diversité.[103]

> **Imposer aux entreprises spécialisées dans l'IA de mener des analyses d'impact internes afin d'identifier d'éventuelles inexactitudes et représentations erronées dans différentes langues avant le déploiement des systèmes d'IA, et les réexaminer en permanence.** Cela suppose également de vérifier la présence d'inexactitudes et de représentations erronées dans le cadre d'audits indépendants (voir chapitre 4, section 2.3) et de permettre aux chercheurs de procéder à un examen approfondi (voir chapitre 4, section 4.2).

E. PROTECTION DES DONNÉES ET DE LA VIE PRIVÉE

Bien que les entreprises affirment prioriser la protection de la vie privée des utilisateurs et la préservation de points de données spécifiques, cet engagement ne s'étend généralement pas aux informations tirées des données comportementales, qui sont la base du capitalisme de surveillance. Dans certains cas, les entreprises recourent délibérément à des stratégies obscures pour empêcher les utilisateurs de bénéficier de garanties solides en matière de protection de la vie privée pour leurs données comportementales. Ces pratiques constituent une entrave majeure à l'exercice effectif des droits des utilisateurs.

Cet engagement limité en faveur de la protection de la vie privée est aussi monnaie courante pour les systèmes d'IA générative, qui recueillent généralement un vaste ensemble de données personnelles et conversationnelles, allant des informations de compte et de communication aux données de connexion et d'utilisation, données pouvant être partagées avec des tiers. Les utilisateurs ne sont pas conscients des informations qu'ils partagent en utilisant l'IA pour en savoir plus sur une situation financière ou médicale dans laquelle ils se trouvent. Ce problème est particulièrement préoccupant pour les journalistes qui pourraient avoir besoin de protéger des informations ou des sources sensibles[104].

Les conversations avec les systèmes interactifs de l'IA sont souvent consultées par les formateurs de l'IA afin d'améliorer les performances du système et à des fins de recherche. Cependant, les droits des utilisateurs en ce qui concerne leurs données et le traitement de leurs entrées et messages ne sont généralement pas clairement communiqués. De plus, bien que les utilisateurs de certains pays aient la possibilité de demander la suppression de leurs données dans les systèmes d'IA générative (droit à

103 Un bon exemple en est BLOOM de BigScience - un LLM open source développé avec le soutien des pouvoirs publics (dont un financement du gouvernement français) - capable de générer du texte dans 46 langues et 13 langages de programmation et d'effectuer des tâches sur lesquelles il n'a pas été explicitement entraîné. Tant BLOOM que le modèle sur lequel il a été formé (c'est-à-dire ROOTs, un ensemble de données multilingue de 1,6 To) sont accessibles à l'examen par d'autres professionnels du TAL. Plus d'informations sur BLOOM de BigScience sont disponibles sur : https://bigscience.huggingface.co/blog/bloom (Consulté le 15 décembre 2023).

104 Des préoccupations similaires pourraient également s'appliquer aux militants pour les droits humains et à d'autres parties prenantes, mais elles dépassent le cadre du rapport et le mandat du Forum sur l'information et la démocratie.

l'oubli[105]), ils n'ont souvent pas la possibilité de supprimer des entrées ou des messages spécifiques.[106] Ce manque de clarté et de contrôle est source de confusion et nourrit des inquiétudes de plus en plus vives en matière de confidentialité.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Concevoir et mettre en œuvre des modèles commerciaux éthiques et des systèmes transparents permettant vraiment aux utilisateurs de refuser d'être suivis et, dans la mesure du possible, d'exercer leur droit à l'oubli. Cela implique :**
>
> - l'arrêt de la collecte de données à caractère personnel au cours des interactions de l'utilisateur de l'IA avec le système d'IA ;
> - l'arrêt de la divulgation des données à caractère personnel du sujet IA dans les résultats générés par le système d'IA ; et
> - dans la mesure du possible, l'effacement de toute donnée personnelle concernant le sujet IA du jeu des données d'apprentissage utilisées par le système d'IA.

> **Concernant les données et les interactions des utilisateurs avec un système d'IA, communiquer clairement, de manière facilement accessible et visible :**
>
> - toute opération de traitement des données à des fins d'apprentissage et d'amélioration du modèle ou d'autres utilisations ;
> - la manière dont les données sont stockées ;
> - si les données sont partagées avec des tiers.

> **Expliquer comment l'utilisation et le stockage des données et des interactions des utilisateurs respectent les lois sur la protection des données et la vie privée.** Ce principe est essentiel pour que les utilisateurs puissent prendre des décisions éclairées concernant leur vie privée et l'utilisation de leurs données.

> **Permettre aux utilisateurs de refuser que leurs données entrantes et leurs interactions soient stockées et utilisées pour l'amélioration du système d'IA,** ce qui devrait être le paramètre par défaut.

105 Inscrit à l'article 17 («Droit à l'effacement (« droit à l'oubli »)») du Règlement général sur la protection des données (RGPD) européen.
106 OpenAI (2023). Privacy policy. Disponible sur : https://openai.com/policies/privacy-policy (Consulté le 15 décembre 2023).

RECOMMANDATIONS
AUX **ÉTATS**

> **Imposer aux entreprises et entités spécialisées dans l'IA de faire la preuve qu'elles réduisent les risques liés à l'utilisation de données protégées.**

> **Instaurer un système de dépôt de plaintes concernant l'utilisation de données protégées, en garantissant le droit à l'oubli et en attribuant la charge de la preuve à l'entreprise** (voir chapitre 2, sections 1 et 4.3).

> **Exiger des entreprises et entités spécialisées dans l'IA qu'elles communiquent clairement aux utilisateurs la manière dont leurs données entrantes et d'interaction sont utilisées et stockées, et qu'elles mettent en place un mécanisme permettant aux utilisateurs de refuser facilement le stockage et l'utilisation de leurs données.**

F. SOUTENABILITÉ ET INDEMNISATION DES MÉDIAS

Les données extraites sur Internet pour former les systèmes d'IA comprennent souvent aussi des données produites et détenues par des organisations médiatiques[107]. Actuellement, le manque de transparence sur la provenance des jeux de données ne permet pas de déterminer l'utilisation exacte des contenus médiatiques dans les systèmes d'IA. Par ailleurs, nous ne savons pas dans quelle mesure les systèmes d'IA sont ajustés et ré-entraînés à partir des données fournies par les utilisateurs, y compris les organisations médiatiques.

Pourtant, les programmes d'IA dépendent de données de haute qualité pour former leurs systèmes. Le plus souvent, ce sont les organisations médiatiques, et plus particulièrement les médias du service public, qui possèdent des données dans les langues locales et autochtones. Il est donc dans l'intérêt des entreprises et des entités spécialisées dans l'IA d'assurer la soutenabilité de ces médias afin de pouvoir accéder à des données de haute qualité provenant de multiples groupes culturels et linguistiques. De plus, les entreprises d'IA conçoivent leurs systèmes, et par conséquent leurs profits, sur la base du contenu fourni par les organisations médiatiques, ce qui pose la question du partage équitable des profits et de l'indemnisation.

La législation sur le droit d'auteur et son application à l'IA, telle que le principe de « l'usage raisonnable [108]» américain ou l'exception européenne relative à la la fouille de textes et de données (text and data mining ou TDM en anglais), souffrent d'un flou juridique qui ne permet pas de déterminer dans quelle mesure l'extraction de contenus médiatiques peut être considérée comme une violation du droit d'auteur[109]. Cela vaut également pour d'autres contenus protégés par des droits d'auteur, qui sortent du cadre du présent rapport. Les organisations médiatiques commencent donc à exercer leur droit de retrait (*opt-out*), en indiquant dans leurs conditions de service que leur contenu ne peut être récupéré par les sociétés d'IA sans autorisation[110], et à poursuivre les sociétés d'IA pour l'utilisation illégale de leur matériel[111].

107 News/Media Alliance (2023). White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement And Not a Fair Use. Disponible sur : www.newsmediaalliance.org/generative-ai-white-paper/ (Consulté le 15 décembre 2023).
108 *Ibid*
109 C'est également vrai pour les contenus protégés par le droit d'auteur, même si cette question dépasse le cadre de ce rapport.
110 Voir par exemple conditions générales d'utilisation de New York Times, disponible sur : https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-Service (Consulté le 7 février 2023).
111 Grynbaum, M.M. and Mac, R. (2023). The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work, The New York Times. Disponible sur : www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (Consulté le 7 février 2024).

Les gouvernements ont commencé à adopter des lois pour garantir la contribution des grandes entreprises technologiques à la soutenabilité et à l'indemnisation des médias, comme en témoignent le Code de négociation australien et la Loi canadienne sur les nouvelles en ligne. Les principes pour une compensation équitable ont été élaborés et adoptés lors de la conférence *Big Tech et journalisme - Construire un avenir durable dans les pays du Sud*, en juillet 2023[112]. Des approches similaires pourraient également être applicables aux entreprises d'IA, comme le prévoient les lignes directrices du Conseil de l'Europe sur la mise en œuvre responsable des systèmes d'intelligence artificielle dans le journalisme[113]. Des droits d'indemnisation exécutoires pourraient être prévus sur la base de différents fondements juridiques tels que le droit d'auteur, le droit de la concurrence ou la fiscalité d'intérêt collectif. Ceux-ci doivent respecter les principes énoncés dans les Principes pour une compensation équitable comprenant notamment le soutien au journalisme d'intérêt public, le maintien de la pluralité et de la diversité des organisations médiatiques et la mise en place d'un système fondé sur l'équité, la soutenabilité et la négociation collective.

Dans l'idéal, les médias de service public devraient fournir des données de haute qualité pour soutenir la mise en place d'alternatives aux systèmes d'IA à but lucratif[114]. Les médias d'intérêt public percevraient une compensation pour leur contribution. Les gouvernements devraient donc adopter une législation permettant aux médias d'être indemnisés pour l'utilisation de leur contenu. Si des accords volontaires entre les entreprises d'IA et les organisations médiatiques peuvent constituer un premier pas, ils ne garantissent cependant pas la soutenabilité et la stabilité des organisations médiatiques et risquent de favoriser les plus importants d'entre eux.

RECOMMANDATIONS AUX **ÉTATS**

> **Définir des droits exécutoires pour les organisations médiatiques et les journalistes en ce qui concerne l'utilisation de leur contenu dans l'apprentissage et les applications de l'IA.** Ceux-ci devraient comprendre :
>
> - **Un droit de savoir,** permettant aux médias d'exiger la transparence sur l'utilisation de leur contenu dans les jeux de données d'apprentissage et de peaufinage.
> - **Un droit de retrait (*opt-out*)**, conférant aux médias la liberté de décider si leur contenu peut être utilisé dans les systèmes d'IA. Cela englobe l'utilisation du contenu dans les jeux de données d'apprentissage, le grounding[115], ainsi que l'utilisation des données saisies par les organismes de médias dans les systèmes d'IA lors de l'utilisation du système.

112 Gordon Institute of Business and Science (2023). Géants de la tech et journalisme : principes pour une compensation équitable. Disponible sur : www.gibs.co.za/news-events/news/pages/big-tech-and-journalism-principles.aspx (Consulté le 7 février 2024).

113 Conseil de l'Europe (2023). Lignes directrices sur la mise en œuvre responsable de systèmes d'intelligence artificielle dans le journalisme. Disponible sur : https://rm.coe.int/cdmsi-2023-014-lignes-directrices-sur-la-mise-en-uvre-responsable-de-s/1680adb4c7 (Consulté le 7 février 2024).

114 Le projet Spinoza de Reporters Sans Frontières est un exemple de cette approche. « Projet Spinoza » : RSF et l'Alliance de la presse d'information générale partenaires pour développer un outil d'intelligence artificielle pour les journalistes. Disponible sur : https://rsf.org/fr/projet-spinoza-rsf-et-l-alliance-de-la-presse-d-information-g%C3%A9n%C3%A9rale-partenaires-pour-d%C3%A9velopper (Consulté le 9 février 2024).

115 Le terme «grounding» désigne l'utilisation des résultats de recherche en temps réel comme contexte pour les réponses des systèmes d'IA, tel que défini par News/Media Alliance : News/Media Alliance (2023). White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement And Not a Fair Use. Disponible sur : www.newsmediaalliance.org/generative-ai-white-paper/ (Consulté le : 7 février 2024).

♦ **Le droit à une compensation équitable lorsque le contenu médiatique est utilisé pour l'apprentissage d'un système d'IA**. Pour garantir le respect de ce droit, plusieurs options sont envisageables :

- Mise en place d'une taxe sur les entreprises d'IA qui réserve expressément :
  - des fonds à redistribuer aux médias au niveau national ou régional ; et
  - une taxe universelle pour soutenir les médias à l'échelle internationale. Les ressources pourraient être distribuées par un fonds mondial indépendant.
- Modifier les lois sur la concurrence de façon ciblée afin de corriger les déséquilibres en matière de pouvoir de marché entre les entreprises et entités d'IA et les organisations médiatiques, en permettant à ces dernières de négocier conjointement les prix, les conditions et les modalités en vertu desquels leur contenu peut être utilisé pour former un système d'IA (c'est-à-dire un système de négociation collective pour le paiement direct aux organisations médiatiques).
- Définir clairement le fait que l'utilisation de contenus médiatiques pour les systèmes d'IA constitue une violation des droits d'auteur, à moins qu'une autorisation ne soit accordée et qu'une compensation ne soit versée.
- Étudier la faisabilité de la mise en œuvre d'un système de redevances semblable aux modèles de rémunération utilisés pour les artistes par des plateformes telles que Spotify.[116] Une telle compensation s'appliquerait non seulement lors de la phase d'apprentissage de l'IA, mais tout au long de l'utilisation des modèles ainsi formés.

> Veiller à ce que les fonds récoltés soient distribués aux médias par des mécanismes indépendants, transparents et responsables, dans le respect de l'indépendance et de l'équité éditoriales et favorisant le journalisme d'intérêt public, la pluralité, la diversité et la soutenabilité des organisations médiatiques.

> À terme, mettre en place des mécanismes de collaboration entre les médias de service public et les alternatives publiques aux systèmes d'IA à but lucratif afin de rémunérer équitablement les médias de service public pour la production de données et de contenus, tout en favorisant le développement d'alternatives publiques aux systèmes d'IA à but lucratif.

1.2. ÉTIQUETAGE HUMAIN POUR L'APPRENTISSAGE DE L'IA

L'étiquetage des données par l'homme est indispensable à l'apprentissage de divers systèmes d'IA, y compris pour les classificateurs et les modèles génératifs. Dans le cadre de cette procédure, les humains évaluent la qualité des réponses des systèmes, générant ainsi des données utilisées ensuite comme retour d'information pour ajuster les performances de l'IA en termes de précision opérationnelle et de normes sociétales. Cet ajustement est avant tout crucial pour les systèmes d'IA générative, car il permet de s'assurer du respect des valeurs humaines et des normes éthiques. Ce processus repose sur des lignes directrices exhaustives et des méthodes éprouvées, essentielles au maintien de la qualité et de l'intégrité de l'apprentissage des modèles.[117]

116 Spotify (2023). Royalties. Disponible sur : https://support.spotify.com/us/artists/article/royalties/ (Consulté le 7 février 2024).
117 OpenAI's (2023). ChatGPT interaction on 5 December 2023.

Cependant, les lignes directrices et les principes spécifiques appliqués dans le cadre de ce processus ne sont pas toujours transparents pour le grand public. Ce flou s'étend également à la logique qui sous-tend certaines décisions, aux critères exacts utilisés, ainsi qu'aux valeurs et aux motivations qui animent les évaluateurs humains impliqués. Il est pourtant indispensable que ces questions soient bien cernées pour garantir la fiabilité et la responsabilité du système d'IA, car elles influencent directement la manière dont le modèle apprend et évolue en interaction avec l'homme.

En raison de l'influence significative que ces décisions peuvent avoir en matière de lutte contre les biais, contre la mésinformation et la désinformation, la haine, la violence graphique et d'autres contenus préjudiciables susceptibles d'apparaître dans les résultats produits par l'IA, les entreprises et entités spécialisées dans l'IA sont appelées à encourager l'ouverture et la transparence des méthodes d'apprentissage (ainsi que le respect de la législation du travail). Cette transparence est essentielle pour attester de leur engagement à s'attaquer efficacement à ces questions capitales. En outre, les entreprises et entités spécialisées dans l'IA devraient associer activement les différentes parties prenantes impliquées au sein du processus de définition de ces lignes directrices et bonnes pratiques, et veiller à ce que le processus d'apprentissage de l'IA reflète et restitue fidèlement la diversité du langage humain et des expressions culturelles. Pour garantir la prise en compte de la diversité des points de vue et des préoccupations dans le perfectionnement des systèmes d'IA, une participation inclusive et une approche soucieuse des spécificités culturelles sont essentielles, car elles permettent une meilleure adéquation de ces systèmes avec l'intérêt collectif.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Améliorer la transparence des procédures d'étiquetage par l'homme dans l'objectif de rendre accessibles au public les politiques, les lignes directrices, les meilleures pratiques et les procédures qui guident les évaluateurs humains au cours du processus d'apprentissage.** Ces informations doivent inclure les définitions utilisées (par exemple, le discours préjudiciable), les domaines/sujets examinés et les critères utilisés pour accepter ou rejeter des résultats spécifiques.

> **Mettre en place des processus participatifs et inclusifs prévoyant au minimum une participation équitable, durable et substantielle de chercheurs indépendants et de la société civile à l'élaboration de politiques, de lignes directrices, de bonnes pratiques et de procédures guidant l'étiquetage et l'ajustement par l'homme au cours du processus d'apprentissage.** Ces processus pourraient impliquer plusieurs partenaires, dont les utilisateurs, les chercheurs et les autorités compétentes. Il faut en particulier que les experts de la société civile soient rémunérés équitablement pour leur contribution et leur expertise (voir chapitre 4, section 3.2).

> **Prendre en compte les spécificités des cultures dans toute leur diversité ainsi que les nuances linguistiques des différentes communautés, en particulier les plus marginalisées historiquement, dans la formulation des lignes directrices et des meilleures pratiques pour les évaluateurs humains.** À cette fin, il est nécessaire que l'étiquetage par l'homme tienne compte des sensibilités culturelles, des variations linguistiques et des styles de communication propres aux différentes communautés à travers le monde.

> **Mettre en œuvre un système à plusieurs niveaux pour les processus d'étiquetage par l'homme, en y intégrant la participation de la société civile.** Ce système fonctionnerait comme suit :

I. Le contenu est d'abord examiné par des évaluateurs humains internes à l'entreprise ou à l'entité d'IA. Leurs commentaires sont utilisés pour former et ajuster le modèle d'IA.

II. Un panel judicieusement constitué de représentants des communautés, notamment des communautés historiquement marginalisées et minoritaires, est invité à examiner un échantillon du contenu évalué par les équipes internes, en vue d'identifier les contenus potentiellement préjudiciables à leur communauté spécifique. Leurs commentaires sont ensuite utilisés pour ajuster le modèle d'IA et améliorer sa compréhension des différents contextes et sensibilités. Il est important que les membres de ces communautés soient rémunérés équitablement pour leur contribution et leur expertise.

III. Un mécanisme simple d'utilisation est mis en œuvre au sein de l'interface du système d'IA pour faciliter les signalements d'éventuels problèmes ou de préoccupations concernant les résultats de l'IA (voir section 2.3).

RECOMMANDATIONS AUX **ÉTATS**

> **Exiger des entreprises et des entités spécialisées dans l'IA qu'elles améliorent la transparence des processus d'étiquetage par l'homme en rendant publiques les politiques, les lignes directrices, les meilleures pratiques et les procédures qui orientent les évaluateurs humains au cours du processus d'apprentissage.**

1.3. MODÉRATION DU CONTENU ET SYSTÈMES DE RÉFÉRENCEMENT

Les systèmes de modération et de référencement des contenus servent de garde-fous, car ils déterminent quelles informations sont diffusées auprès des utilisateurs, et celles qui sont éliminées par filtrage. Ils sont utilisés pour évaluer la pertinence du contenu, sa véracité et son respect des normes de la plateforme, et informer les décisions de déclassement ou de suppression.

À titre d'exemple, une étude portant sur la présence de préjugés raciaux dans les modèles de détection automatique des discours de haine a révélé que ces classificateurs, en partie en raison de leur incapacité à comprendre les nuances contextuelles, avaient tendance à sur-modérer les personnes noires. Plus précisément, les modèles étaient 1,5 fois plus susceptibles de signaler comme offensants ou haineux des tweets rédigés par des utilisateurs s'identifiant eux-mêmes comme personnes noires. En outre, les tweets rédigés en anglais afro-américain se sont révélés « deux fois plus susceptibles » d'être considérés comme « offensants » ou « abusifs ».[118] L'apprentissage des classificateurs à l'aide de données biaisées peut aggraver ces problèmes, entraînant des résultats d'IA faussés ou biaisés.

118 Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N. and Allen, P. (2019). The Risk of Racial Bias in Hate Speech Detection, Association for Computational Linguistics, pp.1668–1678. Disponible sur : https://aclanthology.org/P19-1163.pdf (Consulté le 7 février 2024).

La conception, l'apprentissage et le fonctionnement des classificateurs de modération de contenu et des systèmes de référencement de contenu ont des incidences fondamentales sur la liberté d'expression. Pourtant, les principes qui régissent leur conception sont souvent opaques, ce qui soulève des inquiétudes quant à leur équité et à la possibilité d'encoder des biais systémiques dans les systèmes d'IA.

Pour répondre efficacement à ces défis, les entreprises et entités spécialisées dans l'IA ont tout intérêt à adopter une approche multidimensionnelle et transparente à tous les stades du développement de l'IA, et à garantir la participation active des organisations de la société civile (OSC) et des autres parties prenantes concernés par l'amélioration des systèmes utilisés pour la modération et le référencement des contenus.

En parallèle, les États ou les organismes de régulation sont invités à organiser un processus multipartite afin de définir les contenus préjudiciables et de garantir des systèmes de modération des contenus cohérents et efficaces au sein de leurs juridictions, conformément au droit international des droits humains. À plus long terme, les États pourraient également envisager de constituer des jeux de données publiques d'apprentissage pour les classificateurs et les systèmes de référencement (voir chapitre 3, section 1.5).

RECOMMANDATIONS AUX **ÉTATS**

> **Élaborer une définition commune des différents types de contenus préjudiciables entre les différentes entreprises et entités qui exploitent des classificateurs.** Cette définition unifiée servirait de cadre d'orientation pour l'élaboration et la mise en œuvre de systèmes de modération de contenu, assurant ainsi l'homogénéité et la cohérence du traitement des contenus préjudiciables en ligne, tout au moins au sein d'une même juridiction.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Améliorer la transparence des systèmes de référencement et de modération des contenus. À cette fin, les entreprise et entités pourraient notamment [119]:**

- Expliquer la méthodologie utilisée pour créer les jeux de données sur lesquels les systèmes sont formés, ainsi que le fonctionnement du processus d'annotation.
- Fournir un compte rendu détaillé du mode de sélection des contenus pour les jeux d'apprentissage et du mode de sélection des annotateurs chargés d'étiqueter ces contenus.
- Expliquer le rôle que jouent les signalements de contenus préjudiciables par les utilisateurs dans la constitution des jeux d'apprentissage.

119 GPAI (2023). Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report, Global Partnership on AI. Disponible sur : https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Consulté le 5 décembre 2023).

- Publier régulièrement des données complètes sur les performances des systèmes d'IA, au moyen de mesures standard tenant compte à la fois des faux négatifs et des faux positifs.

> **Instaurer un régime d'essai** pour les systèmes de modération et de référencement des contenus. Celui-ci devrait comporter divers scénarios et intrants afin d'évaluer les performances des systèmes pour toute une série de contenus et de contextes culturels, en veillant à ce que ces contenus soient évalués en fonction de leur contexte spécifique.[120]

> **Prévoir des possibilités de retour d'information et de responsabilisation du public** (voir chapitre 4, section 3).

> **Dans le cadre de la sélection des annotateurs des systèmes de modération, de classement et de réévaluation des contenus, opter pour une approche communautaire incluant des experts linguistiques, des représentants de la société civile, des experts locaux, des défenseurs du patrimoine et de la préservation des langues, des linguistes, des experts en droits humains et des membres de communautés représentant des communautés historiquement marginalisées et des minorités. Ceci suppose :[121]**

- La création de jeux d'apprentissage au niveau local en collaboration avec divers membres de la communauté. Le recours à des « jurys de citoyens [122]» peut contribuer à cette démarche.
- La priorisation des annotateurs visés par les préjudices spécifiques, pour garantir une représentation et une compréhension exactes.
- Le traitement des jeux de données annotées comme une forme de « jurisprudence » informant les décisions des systèmes d'IA et leur permettant d'apprendre des définitions nuancées des contenus préjudiciables au-delà des définitions textuelles.

> **Établir un protocole d'annotation en deux étapes pour classer les contenus préjudiciables sur les plateformes de réseaux sociaux. Dans un premier temps, les annotateurs classent les contenus sous les étiquettes « supprimé », « déclassé », « intact » ou « surclassé ». Ces données sont utilisées pour former un système de modération et de référencement du contenu pour des catégories de contenu particulières. Dans un deuxième temps, les annotateurs classent les paires d'éléments de contenu en fonction de leur nocivité. Ces données sont ensuite utilisées pour former un évaluateur qui délivre une note de nocivité continue pour chaque élément.[123]**

120 Cambridge Consultants (2019). Use of AI in online content moderation, 2019 Report produced on behalf of Ofcom. Disponible sur : www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Consulté le 5 décembre 2023).

121 GPAI (2023). Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report, Global Partnership on AI. Disponible sur : https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Consulté le 5 décembre 2023).

122 Le concept de «jurys citoyens» est emprunté aux systèmes juridiques dans lesquels un groupe de citoyens évalue collectivement l'impact ou le préjudice de contenus spécifiques, de la même manière dont l'on rend le jugement dans les procès en diffamation. Il s'agit également d'une méthode utilisée dans les processus participatifs, reposant sur un système de tirage au sort. Comme expliqué dans : GPAI (2023). Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report, Global Partnership on AI. Disponible sur : https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Consulté le 5 décembre 2023).

123 *Ibid*

> **Former les systèmes de modération et de référencement des contenus à l'aide d' « étiquettes souples » afin d'améliorer la précision et la confiance dans les décisions.** Cela implique le recours à des distributions cibles reflétant la variété des opinions des annotateurs plutôt qu'à des catégories uniques et strictes. Ce processus permet au système d'IA d'apprendre la variabilité des jugements des annotateurs et d'approfondir sa compréhension des nuances du contenu. De plus, en mesurant l'ampleur du désaccord entre les annotateurs sur la catégorisation d'un élément spécifique, un système d'IA peut déterminer le degré de confiance qu'il peut accorder à une décision de modération, assurant ainsi des résultats plus fiables.[124]

> **Intégrer des mesures de désaccord dans les algorithmes de déclassement pour une modération plus nuancée et plus efficace.** Un déclassement moins agressif des éléments susceptibles de susciter un désaccord important garantit une approche plus équilibrée qui tient compte de la diversité des points de vue et des opinions.[125]

> **Ouvrir le marché des classificateurs de contenu préjudiciable afin d'encourager la concurrence, tout en offrant la possibilité de choisir entre différents classificateurs.**

> **Lorsque cela est possible, utiliser des jeux de données d'apprentissage externes pour une catégorie donnée de contenu préjudiciable, créés dans le domaine public, plutôt que de créer un jeu propriétaire et hermétique [126]** (voir chapitre 3, section 1.5).

1.4. OBJECTIFS D'OPTIMISATION

Il est primordial de comprendre et de sélectionner avec soin les objectifs d'optimisation dans le développement des systèmes d'IA, et ce afin de préserver l'intégrité de notre écosystème d'information numérique. Ces objectifs ont valeur de boussole, orientant les algorithmes d'IA pour déterminer ce qui constitue une réussite de leurs opérations.

Dès leur origine, les systèmes d'IA, et plus particulièrement les moteurs de recommandation, ont donné la priorité à l'engagement, défini par certains experts comme « un ensemble de comportements de l'utilisateur, générés dans le cadre normal de l'interaction avec la plateforme, dont on estime qu'ils sont en corrélation avec la valeur pour l'utilisateur, la plateforme, ou d'autres parties prenantes ».[127] Cependant, l'engagement n'est pas toujours synonyme de véritable valeur personnelle ou sociale. Ainsi, les recherches suggèrent que les contenus les plus extrêmes ou les plus chargés émotionnellement ont tendance à susciter davantage d'interaction.[128] Par conséquent, le fait de privilégier l'engagement peut conduire à la promotion de contenus préjudiciables tels que la mésinformation, les discours de haine, voire la violence ethnique, ce qui peut avoir de lourdes répercussions sur les discours et les comportements sociétaux, en particulier dans des situations à forts enjeux telles que la proximité d'une élection, d'un conflit ou d'une pandémie.[129] Cette problématique est particulièrement préoccupante dans un espace numérique où la création de comptes bots et de contenus synthétiques s'avère de plus

124 *Ibid*
125 *Ibid*
126 *Ibid*
127 Bengani, P., Stray, J., & Thorburn, L. (2022). Blog Post: What's Right and What's Wrong with Optimizing for Engagement, Center for Human-Compatible AI at UC Berkeley. Disponible sur : https://humancompatible.ai/news/2022/05/02/blog-post-whats-right-and-whats-wrong-with-optimizing-for-engagement/ (Consulté le 5 décembre 2023).
128 *Ibid*
129 Amnesty International (2022). Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – Report. Disponible sur : www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ (Consulté le 5 décembre 2023).

en plus facile, ce qui permet à ce type d'acteurs de manipuler l'espace informationnel et de saper la confiance. En plus de mesures alternatives de l'engagement, il faudra également, sur le long terme, des méthodes pour améliorer la visibilité du contenu authentifié afin de renforcer la confiance dans l'espace informationnel en ligne et de réduire les risques de pollution de l'espace par les bots des réseaux sociaux.

Bien que des appels aient été lancés pour trouver des alternatives aux systèmes de référencement basés sur l'engagement, tels que les flux chronologiques inversés, ces solutions présentent elles aussi leurs propres limites et ne favorisent pas nécessairement des résultats sociétaux positifs dans l'espace informationnel, ni la confiance dans cet espace. En conséquence, les opérateurs de plateformes ont intérêt à développer et à expérimenter des mesures alternatives pour les systèmes de recommandation et de modération de contenu, qui contribuent à un espace informationnel démocratique, inclusif et digne de confiance.[130] Cette approche encourage le développement et le déploiement de systèmes d'IA plus éthiques (voir chapitre 3, section 1.5).

En parallèle, les chercheurs sont invités à étudier l'impact des mesures d'engagement communément utilisées et des alternatives possibles sur la diffusion de contenus illégaux et légaux mais préjudiciables, parmi lesquels la mésinformation, les théories complotistes, les discours de haine et le harcèlement. En outre, les chercheurs devraient se pencher sur l'impact des systèmes de recommandation sur l'accès à des informations diverses et fiables, sur les questions de débat public et sur les situations à forts enjeux telles que les élections, les conflits, les guerres et les pandémies. Pour y parvenir, il faut permettre à des chercheurs extérieurs de mener des évaluations expérimentales sur les plateformes de réseaux sociaux (voir chapitre 4, section 4.3).[131]

Enfin, la configuration actuelle des algorithmes et des plateformes d'IA favorise l'engagement sur des éléments de contenu spécifiques et accroît leur visibilité, au lieu d'encourager le dialogue constructif et l'échange qui sont fondamentaux pour une société démocratique. À terme, les entreprises et entités d'IA ainsi que les entreprises gestionnaires de plateformes devraient tester et mettre au point d'autres mécanismes d'engagement en ligne.

130 Bengani, P., Stray, J., & Thorburn, L. (2022). Blog Post: What's Right and What's Wrong with Optimizing for Engagement, Center for Human-Compatible AI at UC Berkeley. Disponible sur : https://humancompatible.ai/news/2022/05/02/blog-post-whats-right-and-whats-wrong-with-optimizing-for-engagement/ (Consulté le 5 décembre 2023).

131 Bengani, P., Stray, J., & Thorburn, L. (2022). Blog Post: How to Measure the Effects of Recommenders. Understanding Recommenders, Center for Human-Compatible AI at UC Berkeley. Disponible sur : https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57 (Consulté le 5 décembre 2023).

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Faire face aux potentiels effets négatifs de l'optimisation de l'engagement sur le discours sociétal, le comportement et la démocratie, en particulier dans les situations à forts enjeux.** Cela implique la mise en œuvre d'une série d'étapes déterminantes :

- **Procéder à une évaluation complète des menaces**. Les objectifs sont multiples :
  - **Analyser l'impact des indicateurs d'engagement** (likes, partages, commentaires, taux de clics et temps passé sur le contenu), y compris par des tests A/B, **sur l'amplification des contenus préjudiciables liés aux questions sociales, aux élections, aux conflits et à la politique.**
  - **Identifier les indicateurs d'engagement les plus problématiques** (c'est-à-dire susceptibles de promouvoir des contenus préjudiciables tels que la mésinformation et la désinformation, les discours de haine ou les contenus incitant à la violence) par l'analyse des tendances des données).
- **Élaborer un plan d'urgence pour ajuster les indicateurs d'engagement problématiques dans les situations où ils pourraient conduire à des violations des droits humains, à de la violence ou à des préjudices physiques. Ce plan serait crucial dans des contextes sensibles tels que les élections, les conflits et les urgences sanitaires.** Ce plan pourrait inclure :
  - Des indicateurs alternatifs à substituer ou en remplacement d'indicateurs identifiés comme problématiques.
  - Des stratégies de réduction du poids des indicateurs problématiques dans les algorithmes de référencement.
- **Collaborer avec un groupe diversifié de partenaires incluant des chercheurs, des représentants d'organisations de la société civile, des experts en désinformation et des autorités compétentes, afin de solliciter leurs commentaires et d'intégrer leurs points de vue dans l'évaluation des menaces et le plan d'urgence.**
- **Tester les changements proposés dans le plan d'urgence dans un environnement contrôlé.** Cette démarche peut consister à effectuer des tests A/B ou à appliquer les changements à des groupes démographiques ou géographiques spécifiques.
- **Mettre en place un système de contrôle et d'amélioration continue, garantissant la transparence et un accès simplifié pour les utilisateurs et les tiers de confiance, afin qu'ils puissent faire part de leurs commentaires.**
- **Rendre compte régulièrement des résultats et des ajustements effectués.** Cette étape est essentielle pour responsabiliser les acteurs de l'IA et leur permettre de mieux comprendre l'influence des différents paramètres d'optimisation sur des questions telles que la mésinformation, la haine en ligne, le harcèlement et la violence à l'encontre des communautés marginalisées, ainsi que les stratégies susceptibles de minimiser ces menaces.

> **Permettre aux utilisateurs d'opter facilement et intuitivement pour des systèmes de recommandation alternatifs ne privilégiant pas l'engagement mais des résultats individuels et sociétaux positifs tels que des informations fiables, un contenu fédérateur ou une diversité d'information.**

> **Se livrer à des expérimentations pour évaluer les effets d'autres mesures de l'engagement,** par exemple :

- **Classification par les utilisateurs** suivant des critères tels que la fiabilité du contenu, le caractère recommandable, la clarté et la pédagogie, l'importance et la faisabilité, l'adéquation à un large public, la résilience contre les retours négatifs, l'inclusion et la diversité, et la promotion de bonnes pratiques.[132]
- **Proposer des contenus consensuels ou fédérateurs,** c'est-à-dire approuvés par une diversité d'acteurs.[133] Cela suppose un processus inclusif et transparent pour définir un système d'approbation du contenu.
- **Prioriser un « engagement à long terme » (c.-à-d. sa durée) plutôt qu'un «engagement à court terme» (c.-à-d. son intensité) comme critère dans les systèmes de recommandation.** Des recherches récentes indiquent que le fait d'optimiser l'intensité de l'engagement d'un utilisateur avec un contenu au cours d'une seule session plutôt que le temps passé sur la plateforme sur une période prolongée peut conduire à l'amplification d'un contenu polarisant. Cependant, lorsque l'accent est mis sur le maintien de l'engagement de l'utilisateur dans le temps, les systèmes de recommandation sont plus susceptibles de favoriser des contenus équilibrés et moins clivants.[134]
- **Privilégier les cycles d'engagement plus longs plutôt que les cycles à court terme, qui favorisent les contenus récents, en tenant également compte des contenus plus anciens, à l'exception des contenus et des actualités de dernière minute dont la fiabilité a été certifiée (par exemple par la Journalism Trust Initiative).** En matière de recommandation de contenu, les plateformes diffèrent considérablement quant à leur tendance à privilégier les contenus et événements récents. Des recherches récentes indiquent que les plateformes dont le cycle de vie de l'engagement est plus long gagnent à disposer de plus de temps pour modérer le contenu avant qu'il n'atteigne son pic d'engagement. À l'inverse, les plateformes qui privilégient les actualités et les événements récents ont souvent du mal à modérer efficacement les contenus dans des délais aussi courts. Ainsi, les plateformes dont les périodes d'engagement sont plus longues recommandent généralement des contenus moins polarisants que celles qui suivent des cycles d'engagement plus courts.[135]

> **Adopter des stratégies de conception éprouvées pour réduire la diffusion de contenus préjudiciables.** Les développeurs et déployeurs d'IA sont notamment invités à :

- **Réduire l'optimisation de l'engagement pour les contenus sensibles, notamment en matière de politique et de santé.**[136] Cette mesure qui a fait ses preuves permet de réduire la prolifération de la mésinformation et des contenus clivants.

132 Tournesol Project (Consulté le 7 février 2024).
133 Le contenu fédérateur est «le contenu qui obtient l'approbation (ou génère un engagement positif) auprès de groupes divers de personnes.»Tel que défini dans : Stray, J., Iyer, R., and Puig Larrauri, H. (2023). *The Algorithmic Management of Polarization and Violence on Social Media*, Knight First Amendment Institute at Columbia University. Disponible sur : https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media (Consulté le 15 décembre 2023).
134 Edelson, L., Haugen, F., and McCoy, D. (2023). Demystifying Social Media Feeds: A Taxonomy and Transparency for Algorithmic Feed Systems Designs. Draft manuscript.
135 *Ibid*
136 Stepanov, A. and Gupta, A. (2021). Reducing Political Content in News Feed, Meta. Disponible sur :https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/ (Consulté le 8 février 2024).

- **Renforcer la place des paramètres d'évaluation de la crédibilité et de la qualité des éditeurs de presse et de leurs reportages dans les algorithmes de classement.** Ce type d'initiative permet de garantir que les articles faisant autorité et documentés sont davantage mis en avant, en particulier lorsqu'une échéance électorale approche.[137] La Journalism Trust Initiative propose une solution potentielle en mettant en œuvre un mécanisme international de récompense des pratiques journalistiques éthiques. Conçue comme une norme ISO, cette initiative pourrait être intégrée dans les algorithmes afin de récompenser et de promouvoir le journalisme de qualité.[138] La collaboration avec les instances de gestion des élections peut améliorer l'accès à un contenu électoral fiable en période électorale.[139]

> **Se livrer à des expérimentations visant à modifier les fondements de l'engagement en ligne pour favoriser un débat constructif et un climat de confiance.** Pour cela, il est possible de :

- Renforcer la confiance dans le contenu en ligne en adoptant un mécanisme d'authentification des auteurs s'appuyant sur des techniques cryptographiques préservant la vie privée, telles que les preuves à divulgation nulle de connaissance (zero-knowledge proofs).[140] Ce type de système ne doit cependant pas être utilisé pour renforcer la censure.
- Diffuser à grande échelle les contenus authentifiés d'auteurs ayant choisi d'être vérifiés.
- Fournir aux utilisateurs un parcours de provenance du contenu détaillant ses origines et le statut d'authentification de ses auteurs, améliorant ainsi la transparence et la crédibilité.
- Adopter des systèmes de classement alternatifs visant à créer un dialogue constructif, des compromis et des échanges.

RECOMMANDATIONS AUX **ÉTATS**

> **Instaurer un mécanisme garantissant aux chercheurs indépendants agréés et aux auditeurs externes la possibilité de mener des expériences sur les plateformes[141], telles que des tests A/B, pour mettre en évidence les relations de cause à effet entre la conception et les paramètres des algorithmes et les résultats négatifs tels que la diffusion de contenus préjudiciables et polarisants** (voir chapitre 4, section 4.3).[142]

137 Lyons, K. (2020). Facebook rolls back 'nicer' News Feed that boosted mainstream publishers, The Verge. Disponible sur : www.theverge.com/2020/12/17/22180259/facebook-news-feed-change-post-election-publishers-misinformation (Consulté le 8 février 2024).
138 Journalism Trust Initiative. Disponible sur : www.journalismtrustinitiative.org/ (Consulté le 7 février 2024).
139 Forum sur l'information et la démocratie (2024). Protéger les élections démocratiques par la sauvegarde de l'intégrité de l'information, Forum sur l'information et la démocratie, International IDEA, Democracy Reporting International. Disponible sur : https://informationdemocracy.org/fr/2024/01/30/les-entreprises-technologiques-et-les-gouvernements-sont-appeles-a-lutter-durgence-contre-les-menaces-numeriques-qui-pesent-sur-les-elections/ (Consulté le 8 février 2024).
140 Zero-Knowledge Proof sont des méthodes permettant de vérifier des affirmations sans divulguer les informations réelles, telles que définies dans Aad, I. (2023). Zero-Knowledge Proof. In: Mulder, V., Mermoud, A., Lenders, V., Tellenbach, B. (eds) *Trends in Data Protection and Encryption Technologies* . Springer, Cham. Disponible sur : https://doi.org/10.1007/978-3-031-33386-6_6 (Consulté le 8 février 2024).
141 De telles expériences doivent respecter la confidentialité des données et inclure l'envoi d'une notification aux utilisateurs, comme discuté dans le chapitre 4, section 4.3.
142 Bengani, P., Stray, J., & Thorburn, L. (2022). Blog Post: How to Measure the Effects of Recommenders. Understanding Recommenders, Center for Human-Compatible AI at UC Berkeley. Disponible sur : https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57 (Consulté le 5 décembre 2023).

> **Responsabiliser les plateformes pour qu'elles réduisent la portée des paramètres d'optimisation reconnues pour entraîner des effets négatifs sur l'individu et la société.**

> **Imposer aux entreprises et entités spécialisées dans l'IA de permettre aux utilisateurs de choisir des systèmes de recommandation alternatifs n'optimisant pas l'engagement, mais favorisant des résultats individuels et sociétaux positifs.**

1.5. AUTHENTICITÉ ET PROVENANCE DES CONTENUS

Le développement fulgurant de la génération de contenu par l'IA et la rapidité de sa diffusion compliquent la distinction entre le contenu généré par l'IA et le contenu authentique. Ainsi, la réflexion sur les solutions s'oriente vers la mise au point de mécanismes sophistiqués permettant de garantir l'authenticité et la provenance des contenus.

L'authenticité et la provenance des contenus sont la garantie qu'un contenu numérique représente fidèlement son origine et n'a pas été manipulé (à des fins malveillantes). Cela suppose que ce contenu ait un cycle de vie traçable, c'est-à-dire que sa création, sa modification et sa distribution aient été enregistrées. Plusieurs initiatives ont été lancées pour développer et documenter des normes relatives à l'authenticité du contenu et aux mécanismes de provenance des médias numériques. Par exemple, la Content Authenticity Initiative (CAI) a développé un outil de hachage cryptographique pour fournir des signatures vérifiables et infalsifiables sur les contenus numériques[143], et permettre aux utilisateurs de consulter les données historiques sur le contenu. Un autre organisme, Project Origin, a été créé pour servir de plateforme de discussion sur la création et l'adoption d'un processus de suivi de la provenance des nouveaux médias, initialement pour les actualités et les contenus d'information, parmi un ensemble de partenaires.[144] Enfin, la Coalition for Content Provenance and Authenticity (C2PA) conjugue les actions de la CAI et du Project Origin.[145] Le filigrane (watermarking ou tatouage numérique), ou l'intégration de signes dans les contenus générés par l'IA, est un autre outil couramment utilisé pour établir l'authenticité d'un contenu.

Plus les technologies se perfectionnent, plus les attaques se font sophistiquées, comme en témoigne l'apparition d'outils susceptibles de supprimer ou de contourner les filigranes. Cela met en évidence la nécessité de concevoir des procédés de tatouage numérique plus fiables et plus résistants à la falsification. Le tableau ci-dessous présente une simulation des types et méthodes d'attaques potentielles, ainsi que des possibilités de défense susceptibles d'être perfectionnées pour constituer des outils et des solutions plus efficaces. Outre l'authenticité du contenu et les mécanismes de provenance, les moyens de défense incluent l'amélioration de la formation des utilisateurs pour les aider à identifier les contenus générés par l'IA, l'évaluation régulière des modèles pour mesurer la fiabilité des résultats, des algorithmes de détection, des critères de diffusion du contenu plus stricts, et d'autres formations à la lutte contre les attaques au sein des systèmes.

143 Content Authenticity Initiative (n.d.). How it works. Disponible sur : https://contentauthenticity.org/how-it-works (Consulté le 7 février 2024).
144 Project Origin (n.d.). Project Origin. Disponible sur : www.originproject.info/about (Consulté le 7 février 2024).
145 Au moment de la rédaction du présent document, Meta, Google et OpenAI viennent d'annoncer leur adhésion à C2PA et la mise en œuvre de ses normes, aux côtés de Microsoft qui figure parmi ses fondateurs.

Tableau 1.1 : Synthèse des attaques potentielles contre l'espace informationnel, des méthodes employées et des défenses possibles

ATTAQUE	MÉTHODES	DÉFENSES
Un utilisateur crée et/ou partage involontairement un contenu fallacieux généré par l'IA	Actualités/informations fabriquées de toutes pièces Hallucinations de l'IA Images générées par l'IA	• Formation des utilisateurs • Provenance du contenu • Mécanismes de détection
Un utilisateur crée intentionnellement un contenu trompeur/fallacieux mais n'a pas l'intention/les capacités de contourner les systèmes/ détecteurs de tatouages numériques	Humour, parodie, etc. Images intimes non consenties Spam ou « fermes de contenu » où l'utilisateur est indifférent au fait qu'une partie du contenu soit détectée comme fausse (que ce soit pour des raisons financières ou politiques)	• Filigrane • Provenance du contenu • Mécanismes de détection
Un utilisateur crée intentionnellement un contenu fallacieux ou trompeur et contourne les détecteurs	Suppression ou contournement des filigranes. Contournement des détecteurs par des méthodes n'ayant pas recours aux filigranes (p. ex. ajout de bruit pour dégrader les pixels ou ajout d'informations aléatoires non pertinentes, etc.) afin de rendre plus difficile la détection des manipulations	• Filigranes inviolables • Mécanismes de détection • Provenance du contenu
Un utilisateur crée intentionnellement un contenu fallacieux ou trompeur et en falsifie la provenance	Modification du contenu réel et falsification des données de provenance ou d'authenticité. (Ajout de faux logos ou étiquettes, attaques cryptographiques contre les signatures, vol de clés de signature) Création de faux contenus et falsification de la provenance. (Présentation d'une fausse image/ scène à une véritable caméra, ajout de faux logos ou étiquettes, attaques cryptographiques, vol de clés de signature)	• Formation des utilisateurs • Filigrane • Provenance du contenu • Système d'amplification du contenu / de non-recommandation
Un utilisateur cherche à provoquer des défaillances dans le modèle	Production de résultats de modèles génératifs spécifiques (p. ex. empoisonnement de jeux de données publiques, introduction de bugs dans des générateurs à code source ouvert ou détournement de jeux de données privés) Création de défaillances dans le modèle de détection (p. ex. en introduisant des bugs dans la création ou la détection de filigranes à code source ouvert)	• Conception de modèles fiables • Formation aux techniques adverses • Évaluation/audit régulier du modèle
Un utilisateur tente de modifier le comportement de la plate-forme (par le piratage, etc.)	Manipulation de la plateforme sans accès interne, p. ex. brigading, bots, etc. Blocage du fonctionnement des détecteurs de la plateforme ou remplacement des détecteurs par des détecteurs défectueux. Ajout d'étiquettes trompeuses ou autres modifications de l'interface utilisateur. Modification du classement ou de la modération des contenus.	• Filigrane • Provenance du contenu • Algorithmes de détection • Amplification du contenu / système de non-recommandation

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Mettre au point des techniques de filigrane et de détection de la provenance fiables et rigoureuses (telles que les filigranes ou l'enregistrement du contenu) pour les contenus générés par leurs systèmes d'IA, et garantir leur accès gratuit au public.**[146]

> **Collaborer avec des chercheurs à l'élaboration et à la mise en œuvre d'outils de filigrane et de détection efficaces.**

> **Mettre en œuvre des outils et des normes reconnus concernant l'authenticité et la provenance des contenus. Pour les systèmes de création et d'édition de contenu, il s'agit notamment de générer des signatures numériques et d'autres métadonnées de provenance.**

RECOMMANDATIONS AUX **ÉTATS**

> **Mettre en place un programme de soutien à la recherche pour améliorer les méthodes de détection des contenus générés par l'IA, comme le filigrane ou l'enregistrement des contenus. Ce programme pourrait inclure l'accès à une infrastructure technique de pointe ainsi que le financement et la collaboration avec des plateformes.** Quant aux travaux de recherche, ils devraient s'appuyer sur la littérature existante et approfondir les questions en suspens et les limites actuelles pour les tâches génératives (c.-à-d. l'examen des modèles de diffusion[147]) et discriminatives, les modèles linguistiques[148] et le faux contenu généré par les réseaux antagonistes génératifs (GAN).[149]

> **Définir des normes complètes sur les exigences en matière de filigrane invisible indiquant la provenance et d'étiquetage visible au moyen d'un processus participatif. Cette démarche suppose :**

- Des obligations de filigrane invisibles indiquant la provenance, à inclure dans chaque système d'IA utilisé.
- Un étiquetage visible du contenu généré par l'IA (voir chapitre 2, sections 3.2 et 4.1).
- Des précisions sur les exigences techniques.

> **Définir des normes d'authenticité et de provenance des contenus, y compris en matière d'authentification des auteurs, via un mécanisme participatif incluant la société civile et la communauté universitaire. Ce mécanisme pourrait reposer sur les processus, les normes et les solutions techniques élaborés par la Coalition for Content Provenance and Authenticity (C2PA), afin d'établir une norme reconnue et d'assurer une cohérence entre les systèmes d'IA et les systèmes d'authentification des auteurs et de provenance des contenus.**

146 La proposition de règlement de l'UE sur l'IA oblige les fournisseurs de systèmes d'IA, y compris les systèmes d'IA à usage général, à concevoir des systèmes de telle façon que les images, textes, contenus audio et vidéo à caractère synthétique fassent l'objet d'un marquage dans un format lisible par machine et que leur nature de contenus générés ou manipulés artificiellement soit détectable. (Article 52)

147 Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M. and Lin, M. (n.d.). A Recipe for Watermarking Diffusion Models, arXiv. Disponible sur : https://arxiv.org/pdf/2303.10137.pdf (Consulté le 7 février 2024).

148 Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. and Goldstein, T. (n.d.), A Watermark for Large Language Models, arXiv. Disponible sur : https://arxiv.org/pdf/2301.10226.pdf (Consulté le 7 février 2024).

149 Yu, N. et al (2022). Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. Disponible sur : https://arxiv.org/abs/2007.08457 (Consulté le 7 février 2024).

> **Imposer aux plateformes de détecter les informations sur la provenance et le contenu généré par l'IA** par les méthodes les plus efficaces actuellement disponibles **et de présenter ces informations aux utilisateurs finaux.**

RECOMMANDATIONS AUX **PLATEFORMES**

> **Appliquer les normes existantes sur la provenance du contenu (telles que celles élaborées par la Coalition for Content Provenance and Authenticity), la détection et le filigrane pour renforcer l'intégrité de l'information en ligne et en améliorer la visibilité pour les utilisateurs. Pour les plateformes, navigateurs et autres applications destinées aux utilisateurs finaux affichant du contenu, il s'agit notamment de vérifier la présence de métadonnées de provenance et de les présenter aux utilisateurs, afin de les aider à identifier le contenu authentifié.**

2. TESTS ET ATTÉNUATION DES RISQUES LIÉS AUX SYSTÈMES D'IA

Comme indiqué précédemment, les systèmes d'IA peuvent involontairement entretenir des biais préjudiciables, enfreindre les cadres juridiques relatifs à la protection de la vie privée et des données, voire menacer les valeurs démocratiques et les droits humains, y compris en favorisant la désinformation et en dégradant la protection de la vie privée. Ces risques ont récemment été pointés par le G7, qui a appelé les entreprises et les entités spécialisées dans l'IA à prendre les mesures qui s'imposent tout au long du cycle de vie de l'IA, y compris lors des étapes de développement, de déploiement et de mise sur le marché, afin d'identifier, d'évaluer et d'atténuer les risques en question.[150]

Afin de tester efficacement la sécurité des nouveaux systèmes d'IA qui seront déployés dans l'espace informationnel et d'appliquer les engagements contenus dans le Code de conduite volontaire international[151] adopté par les dirigeants du G7, les entreprises d'IA devraient généraliser les tests de red-teaming afin de couvrir un plus grand nombre de langues et de contextes culturels, pour tenir compte de l'utilisation de l'IA à l'échelle mondiale. En outre, elles devraient activement collaborer avec des chercheurs extérieurs et des organisations de la société civile susceptibles d'identifier les biais et les problèmes éthiques qui pourraient échapper aux développeurs de l'IA. Pour que cette démarche essentielle ne se limite pas à une simple case à cocher, l'engagement doit être clair, durable, substantiel et rémunéré.

150 Commission Européenne (2023).Code de conduite international pour les systèmes d'IA avancés dans le cadre du processus Hiroshima. Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Consulté le 8 février 2024).

151 *Ibid*

L'évaluation des risques avant la mise sur le marché est également essentielle au respect des valeurs démocratiques et des droits humains. Pour être efficaces, les évaluations des risques doivent se concentrer sur des cas d'utilisation spécifiques et évaluer méthodiquement les risques liés à la modération des contenus, à l'exposition à des contenus illégaux ou préjudiciables, ainsi qu'à la diffusion et à l'impact de la désinformation, particulièrement en matière d'intégrité électorale. En parallèle, elles peuvent identifier et atténuer les risques de manière proactive en amont, et suivre et traiter en continu les lacunes potentielles tout au long du cycle de vie de l'IA. La publication de « cartes de modèles » détaillant les utilisations de l'IA, les stratégies d'atténuation et la description des processus de formation et de tests peut renforcer le contrôle et l'atténuation continus des risques, tout en promouvant la responsabilisation.

Enfin, même après des évaluations rigoureuses préalables au lancement et des contrôles de conformité, les systèmes d'IA nécessitent une vigilance constante pour remédier à l'empoisonnement potentiel des données et des modèles, pour faire face à des résultats imprévus dans le monde réel, pour garantir la conformité avec des réglementations et des normes en constante évolution et pour mettre en place rapidement des mesures correctives en cas de dérives. À cet égard, le contrôle après la mise sur le marché peut protéger efficacement les systèmes d'IA contre l'altération des données et des modèles et garantir une évolution de leur conformité en parallèle avec celle de la réglementation. Il est également essentiel de mettre en place un mécanisme structuré de retour d'information de la part des utilisateurs, des protocoles de résolution rapide des problèmes, une collaboration avec des signaleurs de confiance et des mises à jour régulières des systèmes d'IA.

2.1. RED-TEAMING

Bien que les définitions et les objectifs des activités de red-teaming en matière d'IA soient très variables, il existe un fort consensus sur le fait que les activités de red-teaming sont un élément clé de la gestion des risques liés à l'IA.[152] Ainsi, des documents tels que le décret de la Maison Blanche[153], les principes du G7[154] et les *Emerging Processes for Frontier AI Safety* du gouvernement britannique[155], publiés avant la déclaration de Bletchley Park, ont unanimement cité le red-teaming comme méthode privilégiée de gestion des risques liés à l'IA.

À l'heure où la communauté de l'IA travaille à l'élaboration d'une définition unifiée du red-teaming, il est tout aussi essentiel de concevoir des stratégies ciblées dans le cadre des efforts actuels de red-teaming afin d'identifier les vulnérabilités potentielles des systèmes d'IA générative conçus pour être utilisés dans l'espace informationnel.

Premièrement, les tests de red-teaming devraient dépasser les contextes anglophones et occidentaux afin de réduire le risque de biais et d'utilisation abusive. Deuxièmement, la composition des red teams devrait refléter la diversité des utilisateurs et inclure des compétences variées en la matière. Troisièmement, les membres des red teams devraient collaborer avec des chercheurs extérieurs indépendants. Quatrièmement, les développeurs et déployeurs d'IA sont invités à rendre l'IA plus facile à comprendre en publiant des cartes de modèles d'IA avant de commencer à développer/déployer des systèmes d'IA, et éventuellement à les soumettre à l'évaluation préalable du public. Cinquièmement,

152 Frontier Model Forum (2023). Frontier Model Forum: What is Red Teaming? Disponible sur : www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf (Consulté le 7 février 2024).
153 The White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Disponible sur : www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Consulté le 7 février 2024).
154 Commission Européenne (2023).Code de conduite international pour les systèmes d'IA avancés dans le cadre du processus Hiroshima. Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Consulté le 7 février 2024).
155 UK Government (2023). Emerging Processes for Frontier AI safety. Disponible sur : https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (Consulté le 8 février 2024).

dans le cadre de processus démocratiques ou participatifs, il est tout aussi indispensable de définir les lignes directrices du red-teaming, de garantir la transparence de ces règles et de créer des mécanismes de red-teaming externe et continu des modèles d'IA afin d'identifier les risques en matière de droits et de sécurité.

S'il est essentiel que les entreprises appliquent immédiatement ces mesures, il est également urgent que les organismes internationaux de normalisation définissent des pratiques obligatoires et normalisées de red-teaming pour les systèmes d'IA. La portée des activités des pratiques actuelles de red-teaming pour les systèmes d'IA générative inclut des techniques telles que le piratage pour extraire des données sensibles, l'injection de prompt et la manipulation du système dans le but de générer des résultats préjudiciables tels que des cyber-malwares, des formules de produits chimiques toxiques viraux, et des stratégies d'attaque terroriste. Ces actions présentent des risques importants, notamment en raison de leur potentiel d'exploitation par des tiers malveillants. Si cette lacune réglementaire n'est pas comblée, les initiatives de red-teaming non réglementées pourraient représenter une menace existentielle grave pour la sécurité publique et l'intégrité des systèmes démocratiques.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Étendre le champ d'application des tests de red-teaming afin de couvrir un plus grand nombre de langues et de contextes culturels.** Cette mesure renforcera la capacité des red teams à identifier les vulnérabilités des systèmes d'IA au-delà des contextes anglophones et occidentaux prédominants.

> **Définir des lignes directrices pour le red-teaming d'une manière** participative et inclusive comprenant à minima une participation équitable, durable et substantielle des chercheurs indépendants, de la société civile et des communautés affectées, et **mettre ces lignes directrices à la disposition du public.**

> **Fournir aux autorités compétentes des informations détaillées sur les performances des modèles d'IA lors des tests de red-teaming. Ces informations doivent également s'accompagner d'une description des mesures mises en œuvre pour résoudre les problèmes identifiés et renforcer la sécurité globale du système.**[156] Le degré de détail des informations requises dépendra des risques systémiques que le système d'IA fait peser sur l'espace informationnel (voir section 2.2).

> **Composer les équipes de red teams de façon à ce qu'elles reflètent la diversité des utilisateurs en termes de démographie, de langues et d'origines culturelles. Elles doivent également inclure des représentants des communautés marginalisées et vulnérables, qui sont souvent les plus susceptibles de subir des préjudices.**

> **Veiller à ce que les membres des red teams disposent d'une expertise dans un grand nombre de domaines, y compris :**

- Les droits de l'homme et les droits civils
- L'éthique et les normes journalistiques
- L'intégrité électorale

156 The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Disponible sur : www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Consulté le 7 février 2024).

- La protection des données
- La propriété intellectuelle
- La mésinformation et la désinformation, la haine en ligne, le harcèlement, l'extrémisme et les propos illégaux

> **Encourager la collaboration avec des chercheurs extérieurs dans le cadre des activités de red-teaming afin de combler les lacunes en termes de connaissances et d'éviter la pensée grégaire, tout en soulignant la nécessité d'une rémunération équitable et de pratiques éthiques.**

> **Réaliser des tests de red-teaming tant au niveau du modèle qu'au niveau de l'application** afin d'assurer une évaluation et une atténuation complètes des risques à tous les stades du développement et du déploiement des systèmes d'IA.

> **Investir dans la définition de normes et de mécanismes de red-teaming externe et permanent des modèles de fondation et de leurs applications.** Cette méthode a démontré son efficacité en permettant d'identifier non seulement les risques émergents en matière de sûreté et de sécurité de l'IA, mais aussi des questions essentielles telles que les biais, les discriminations et la protection de la vie privée.[157]

RECOMMANDATIONS AUX **ÉTATS**

> **Soutenir les initiatives collectives visant à définir des lignes directrices complètes pour la réalisation de tests de modèles d'IA par les red-teams.** Ces lignes directrices devraient inclure des procédures et des méthodologies garantissant la rigueur des tests et des évaluations, et aborder des questions telles que les biais, la discrimination, la mésinformation et la désinformation, les discours de haine et d'autres contenus susceptibles de porter atteinte aux droits humains.

> **Imposer aux entreprises et aux entités spécialisées dans l'IA de fournir aux autorités compétentes des informations détaillées sur les performances des modèles d'IA lors des tests de red-teaming, et les tenir responsables de la conformité des tests de red-teaming avec les lignes directrices établies.**

> **Imposer aux entreprises et aux entités spécialisées dans l'IA de rendre publiques leurs lignes directrices en matière de red-teaming.**

2.2. ÉVALUATION DES RISQUES AVANT DÉPLOIEMENT

Bien que le red-teaming contribue à identifier les vulnérabilités des systèmes d'IA, il présente aussi des limites en termes d'évaluation des catégories fluctuantes de préjudices au sein de l'écosystème de l'information, comme la mésinformation et les discours de haine.[158]

157 Mislove, A. (2023). OSTP Blog Post: Red-Teaming Large Language Models to Identify Novel AI Risks, The White House, Office of Science and Technology Policy. Disponible sur : www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/ (Consulté le 7 février 2024).

158 Robertson, D. (2023). *When 'red-teaming' AI isn't enough*, POLITICO. Disponible sur : /www.politico.com/newsletters/digital-future-daily/2023/10/25/when-red-teaming-ai-isnt-enough-00123577 (Consulté le 7 février 2024).

Des mesures supplémentaires, telles que des évaluations complètes des risques liés à l'IA avant le déploiement[159], peuvent faire office de garde-fou dans ce domaine, en garantissant un développement éthique et une évaluation complète des systèmes d'IA influant sur l'espace informationnel avant leur lancement. Ces évaluations permettent également de déterminer le niveau de risque (élevé, moyen ou faible) d'un système et, par conséquent, les réglementations auxquelles les systèmes d'IA doivent se conformer (voir chapitre 4, section 2).

Parmi les risques systémiques que les systèmes d'IA font peser sur l'espace informationnel, figurent notamment :

- Les risques liés à la diffusion de contenus illégaux.
- Les risques affectant l'exercice des droits fondamentaux, notamment la liberté d'expression et d'information, y compris le droit d'accès à l'information, la liberté et le pluralisme des médias, le droit de vote et la participation politique, le droit à la vie privée, la protection des données et le droit à la non-discrimination.
- Les risques ayant une incidence sur les processus démocratiques, le discours civique et les processus électoraux, ainsi que sur d'autres situations à forts enjeux, telles que la santé publique et le maintien de la paix.[160]

Le tableau 1.2 présente une vue d'ensemble des préjudices potentiels susceptibles de résulter de l'utilisation de systèmes d'IA dans l'espace de l'information et de la communication, et pouvant constituer un risque systématique. Il fournit également des indications sur les facteurs à prendre en compte pour évaluer la probabilité qu'un tel préjudice se produise, ainsi que sa gravité potentielle. Ce tableau doit être régulièrement mis à jour en fonction des derniers développements technologiques et des risques qui en découlent.

Tableau 1.2 : Évaluation des risques systémiques des systèmes d'IA pour l'espace informationnel

PRÉJUDICES POTENTIELS À DÉTERMINER DANS LE CADRE DE L'ÉVALUATION DES RISQUES	FACTEURS À PRENDRE EN COMPTE POUR ÉVALUER LA PROBABILITÉ ET LA GRAVITÉ DE LA SURVENUE DES PRÉJUDICES[161]
• Infractions aux droits d'auteur • Violations de la vie privée et de la protection des données • Questions relatives à l'authenticité et à la fiabilité • Création et diffusion de mésinformations et d'hallucinations • Création et diffusion de désinformation et de deepfakes • Création et diffusion de propos illégaux, de discours de haine et de violence • Biais, discrimination et hégémonie culturelle dans les contenus générés par l'IA • Surveillance et exploitation des données • Manipulation, tromperie et usurpation d'identité • Hyperpersonnalisation • Polarisation et escalade des conflits • Censure • Harcèlement	• Objectif recherché par le système d'IA • Capacité du système d'IA à provoquer des préjudices (résultats du red-teaming, qualité des jeux de données, etc.) • Capacité du système d'IA à réagir aux préjudices et à les corriger (mesures d'atténuation des risques, mécanismes de retour d'information et de réclamation, procédures de correction) • Transparence du système d'IA et contrôle externe • Sécurité et résistance du système d'IA contre les utilisations abusives (c.-à-d. cybersécurité) • Accessibilité du système d'IA au grand public • Accessibilité du système d'IA aux acteurs malveillants • Nombre d'utilisateurs réels et potentiels • Utilisation des systèmes d'IA par les acteurs stratégiques de l'espace informationnel (c.-à-d. médias, gouvernement) • Type et quantité de données d'apprentissage • Capacité des systèmes d'IA à agir de manière autonome • Antécédents des préjudices causés par le système d'IA

159 À titre de comparaison, la proposition de règlement de l'UE sur l'IA oblige les déployeurs de systèmes d'IA à haut risque à effectuer une évaluation de l'impact sur les droits fondamentaux avant de mettre un système sur le marché (voir article 29 bis).
160 Inspiré par les risques systémiques cités à l'article 34 du règlement de l'UE sur les services numériques.
161 Voir les critères d'évaluation des risques de préjudices causés par les systèmes d'IA dans la proposition de règlement de l'UE sur l'IA, article 7.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Évaluer les risques systémiques encourus par l'espace informationnel issus de l'utilisation de systèmes d'IA dans la vérification et la modération de contenu, la curation de contenu et la recommandation, le ciblage et la diffusion publicitaires, la création de contenu et la personnalisation de contenu.** À cette fin, il convient de demander l'avis des parties prenantes concernées, notamment les OSC, les chercheurs et les représentants des communautés affectées, et de les rémunérer équitablement pour leur expertise.

- **Ne pas se contenter d'une évaluation théorique des risques, mais prévoir également des cas d'utilisation spécifiques** dans les procédures et les rapports d'évaluation des risques.

> **Mettre au point des « cartes de modèles » pour les systèmes d'IA à haut risque et les mettre à la disposition des régulateurs. Cette mesure devrait exclusivement s'appliquer aux nouveaux systèmes d'IA et aux modifications notables apportées aux systèmes existants. Les cartes de modèles devraient être publiées préalablement à l'élaboration du modèle, évaluées avant leur déploiement et régulièrement mises à jour pour plus de transparence. Ces cartes de modèles devraient préciser :**

- **les mauvaises utilisations potentielles du système d'IA ;**
- **les stratégies d'atténuation ;**
- **les processus de formation et de test prévus, y compris les données d'apprentissage, avant et pendant le développement du système.**

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux entreprises et entités spécialisées dans l'IA de publier des « cartes de modèles » pour les systèmes d'IA à haut risque avant leur développement, et de les mettre à jour régulièrement.**

> **Imposer aux entreprises et entités du secteur de l'IA de procéder à des évaluations obligatoires du risque systémique avant de déployer des systèmes d'IA et d'en présenter les résultats de manière transparente, vérifiable et accessible** par dépôt public (voir chapitre 4, sections 2.1 et 4.1). Une telle approche permet de disposer d'un registre clair et fiable des risques potentiels associés à chaque système d'IA.

2.3. SYSTÈMES DE CONTRÔLE APRÈS DÉPLOIEMENT ET MESURES D'ATTÉNUATION DES RISQUES

Même lorsque des évaluations minutieuses ont été effectuées avant leur lancement, les systèmes d'IA requièrent une vigilance constante. Les tests préalables au lancement, bien qu'essentiels, ne peuvent pas entièrement prédire les complexités et les situations inattendues rencontrées dans les applications du monde réel.

La menace permanente d'empoisonnement des données et des modèles suscite de vives inquiétudes. Des entités malveillantes pourraient manipuler des données ou modifier des modèles d'IA, compromettant ainsi leur intégrité et conduisant à des résultats biaisés ou erronés. En outre, la portée des évaluations des risques avant le lancement peut ne pas couvrir tout le spectre des comportements et des décisions des systèmes d'IA dans le monde réel, ce qui nécessite un contrôle continu pour identifier et réduire les risques émergents.

Ces menaces et limitations potentielles peuvent entraîner des risques imprévus et des conséquences inattendues, d'où l'importance d'une mise en place de systèmes de contrôle robustes après le déploiement incluant des mises à jour régulières pour l'évaluation des risques systémiques ainsi qu'une maintenance périodique. En outre, un contrôle efficace après le déploiement doit être accompagné de mesures d'atténuation des risques et de modération des résultats adéquates. Ces mesures comprennent la mise en place de mécanismes structurés permettant aux utilisateurs et aux signaleurs de confiance de signaler les problèmes et les doutes, des partenariats avec des fact-checkers, ainsi que des protocoles de réponse et de rectification rapides en cas de risques ou de défaillances constatés.

Bien que de tels mécanismes soient déjà requis pour les très grandes plateformes en ligne (VLOP) en Europe, en vertu du règlement sur les services numériques (DSA) et de la loi britannique sur la sécurité en ligne, un mécanisme d'atténuation des risques et de modération des résultats similaire devrait également s'appliquer aux systèmes d'IA générative, permettant d'identifier et de traiter les résultats et contenus indésirables ou préjudiciables et de garantir le respect et la protection des droits.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Adopter des mesures d'atténuation des risques et de modération des résultats afin d'actualiser le système en continu et de répondre aux risques identifiés. Celles-ci doivent comprendre :**
>
> - **Des dispositifs de notification et d'action pour recueillir les commentaires des utilisateurs en privé** (voir chapitre 2, sections 2 et 3.1) **et le retour d'information accessible au public** (chapitre 4, section 3.1).
> - **Une procédure de dépôt de plainte.**
> - **Une collaboration avec des signaleurs de confiance et des fact-checkers.**
> - **Un travail de red-teaming continu pour faire face aux risques identifiés.**

> **Mettre à jour périodiquement l'évaluation des risques systémiques après le déploiement du système d'IA et modifier ce dernier en fonction des nouveaux risques identifiés. Rendre les résultats de ces évaluations accessibles au public.**

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux déployeurs de systèmes d'IA la mise en place de mesures d'atténuation des risques, notamment :**
  - ♦ **Des dispositifs de notification et d'action pour recueillir les commentaires des utilisateurs** (voir chapitre 2, sections 2 et 3.1).
  - ♦ **Une procédure de dépôt de plainte pour contester les résultats générés qui enfreignent les droits existants, y compris des voies de recours claires** (voir chapitre 2, sections 3.1, 3.2 et 4.3).
  - ♦ **Des protocoles de réponse et de résolution rapides en cas d'identification de problèmes.**
  - ♦ **Un travail de red-teaming en continu.**
  - ♦ **Une collaboration structurée avec des signaleurs de confiance et des fact-checkers.**

> **Imposer aux déployeurs de systèmes d'IA des opérations de mise à jour régulières des évaluations du risque systémique, ainsi qu'une maintenance et des mises à jour du système en fonction des résultats de l'évaluation. La fréquence de ces mises à jour est déterminée en fonction du nombre d'utilisateurs du système d'IA et du risque systémique qu'il représente pour l'espace informationnel. Ces évaluations doivent prendre en compte les problèmes identifiés et signalés par les utilisateurs** (voir chapitre 4, section 3.1) **et être rendues publiques.**

CHAPITRE 2 :
RÉGIMES DE RESPONSABILITÉ

INTRODUCTION

Les systèmes d'IA sont susceptibles de produire des résultats erronés, inexacts, partiaux voire discriminatoires. Ils peuvent très concrètement enfreindre les lois existantes en matière de protection de la vie privée, de protection des données, de propriété intellectuelle et de lutte contre la diffamation ou les discours de haine. En outre, ils restent vulnérables aux utilisations malveillantes, qui vont de la manipulation de l'opinion publique à la mésinformation et à la désinformation, en passant par la production de deepfakes et autres contenus falsifiés.

L'intégration de plus en plus marquée des systèmes d'IA dans notre espace informationnel et de communication et la façon dont ils le transforment soulèvent des questions fondamentales. Ces enjeux ont une incidence sur les régimes de responsabilité qui sous-tendent le contrat social entre les développeurs et les déployeurs de systèmes d'IA et la société démocratique.

En premier lieu, la présence d'une multitude d'acteurs dans l'écosystème de l'IA crée un « problème de multiplicité des mains », où la responsabilité est diluée entre de nombreux protagonistes, tout au long de la chaîne de valeur. Puis, l'interaction entre les humains et les systèmes d'IA, approche dite de « l'humain dans la boucle », complexifie davantage l'attribution des responsabilités.[162] L'une des questions clés consiste à savoir dans quelle mesure la prise de décision peut être laissée aux systèmes d'IA. Enfin, l'interaction entre de multiples agents d'IA, en particulier lorsqu'elle se produit très rapidement et à grande échelle, peut produire des résultats imprévisibles et potentiellement préjudiciables. Même après de nombreux tests et ajustements, l'interaction d'agents d'IA individuels, chacun en capacité d'apprentissage et d'amélioration itérative, présente des risques difficilement prévisibles ou maîtrisables.[163]

Ces différents cas de figure mettent tous en question l'attribution de la responsabilité des décisions et des résultats de l'IA. Il est impératif d'aborder ces questions pour fournir aux utilisateurs et aux sujets de l'IA une protection juridique adéquate contre de potentielles atteintes aux droits humains et violations pénales, afin de préserver la confiance dans les technologies de l'IA utilisées pour créer, diffuser et consommer de l'information. Le projet de convention-cadre du Conseil de l'Europe sur l'intelligence artificielle, les droits humains, la démocratie et l'État de droit[164] insistent également sur la nécessité d'établir la responsabilité des violations des droits humains et de prévoir des voies de recours efficaces.

Ce chapitre examine comment les décideurs politiques devraient attribuer les responsabilités tout au long de la chaîne de valeur de l'IA et veiller à ce que les développeurs et déployeurs de systèmes d'IA assument pleinement ces responsabilités. Il souligne ensuite la nécessité d'inverser la charge de la preuve dans les cas de responsabilité impliquant des systèmes d'IA. Enfin, il présente des mesures visant à permettre aux sujets de l'IA de tenir les parties en cause pour responsables des préjudices causés et des dommages infligés.

---

162 Yeung, K. (2019). Responsibility and AI, Council of Europe study. Disponible sur : https://rm.coe.int/responsability-and-ai-en/168097d9c5 (Consulté le 8 février 2024).
163 *Ibid*
164 Conseil de l'Europe (2023). Projet de Convention-cadre sur l'Intelligence Artificielle, les Droits de l'Homme, la Démocratie et l'Etat de Droit. Disponible sur : https://rm.coe.int/cai-2023-28-fr-projet-de-convention-cadre/1680ae19a1 (Consulté le 7 février 2024).

1. ASSUMER LA RESPONSABILITÉ DES RÉSULTATS ET DES DÉCISIONS DE L'IA

1.1. PRINCIPAUX ACTEURS DE LA CHAÎNE DE VALEUR DE L'IA ET LEURS FONCTIONS

L'intégration des systèmes d'IA dans l'espace de l'information et de la communication présente des risques potentiels d'atteintes aux droits humains, comme le souligne la *Taxonomy of Human Rights Risks Connected to Generative AI* (*Classification des risques pour les droits humains liés à l'IA générative*)[165] des Nations unies. Ces atteintes peuvent être à la fois tangibles, comme les risques sanitaires dus à la diffusion de fausses informations médicales, et systémiques, notamment les biais dans les contenus générés par l'IA qui altèrent les modèles sous-jacents et perpétuent les inégalités. Cette intégration remet également en question l'accès à des informations fiables au sein de l'écosystème informationnel, tel qu'il est défini dans les principes du Partenariat pour l'information et la démocratie.[166]

Il est primordial de déterminer les responsabilités en matière d'atténuation de ces risques pour protéger nos démocraties, surtout au regard de la complexité de l'écosystème de développement, de mise en œuvre et d'utilisation de l'IA. Cependant, il n'existe pas de méthode standard pour répartir proportionnellement les responsabilités, ni de mécanismes permettant de garantir que les développeurs et les utilisateurs de systèmes d'IA assument pleinement ces responsabilités.

Les pratiques éthiques volontaires et les codes de conduite du secteur technologique (examinés au chapitre 3, section 1) attestent de la nécessité de prendre très au sérieux cette question de la responsabilité. Ces initiatives manquent toutefois souvent de mécanismes institutionnels de validation externe, de dispositifs d'application et de sanctions sévères, ce qui les rend inaptes à constituer des garde-fous efficaces contre les effets néfastes de l'IA.

Définir clairement ces responsabilités et leur exercice au moyen d'obligations juridiques exécutoires est une nécessité établie dans la Déclaration universelle des droits de l'homme (DUDH), qui exige des États qu'ils protègent les individus et les groupes contre les atteintes aux droits humains. Cette responsabilité implique l'adoption de mesures préventives pour garantir la pleine réalisation des droits humains, en particulier les droits essentiels à la préservation d'un espace mondial de l'information libre, pluriel et diversifié.[167] Ces mesures comprennent la définition d'obligations proactives pour les entités externes, y compris les entreprises et les entités d'IA, qui sont donc tenues d'adhérer à des mandats juridiques garantissant le respect des principes des droits humains.

165 OHCHR (2023). Taxonomy of Human Rights Risks Connected to Generative AI. Disponible sur : www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf (Consulté le 7 février 2024).
166 Forum sur l'information et la démocratie. Partenariat International sur l'Information et la Démocratie. Disponible sur : https://informationdemocracy.org/fr/partenariat-international-information-democratie/ (Consulté le 8 février 2024).
167 Organisation des Nations Unies. La DUDH : Fondement du droit international relatif aux droits de l'homme. Disponible sur : www.un.org/en/about-us/udhr/foundation-of-international-human-rights-law (Consulté le 8 février 2024).

L'attribution d'obligations et de devoirs juridiques exécutoires, profondément ancrée dans les principes de la DUDH (en particulier les droits fondamentaux à la justice et à un recours effectif[168]) est indispensable pour garantir la responsabilité en cas de violations des droits humains par des systèmes d'IA.

RECOMMANDATIONS AUX **ÉTATS**

> En premier lieu, les responsables politiques devraient :

- **Identifier clairement les acteurs clés de l'écosystème de l'IA et leurs responsabilités.**[169] Il doit s'agir à minima des développeurs d'IA, des déployeurs d'IA et des utilisateurs de systèmes d'IA à des fins personnelles, professionnelles ou commerciales. Cette étape est primordiale pour s'assurer que toutes les parties concernées sont conscientes de leur rôle et de leurs responsabilités.
- **Veiller à ce que le niveau de responsabilité attribué corresponde aux risques systémiques que le système d'IA représente pour l'espace informationnel,** ce qui implique que plus les risques sont élevés, plus les exigences à satisfaire sont importantes (voir chapitre 1, section 2.2, et chapitre 4, section 2).
- **Identifier et attribuer la responsabilité des conséquences involontaires et de l'utilisation délibérément erronée ou abusive des systèmes d'IA.** Cette démarche couvre l'ensemble des problèmes potentiels liés au développement et au déploiement de l'IA afin de garantir l'obligation de rendre des comptes.

> Une fois les principaux acteurs de l'IA et leurs responsabilités clairement identifiés, les responsables politiques doivent préciser tout aussi clairement leurs attributions. À cette fin, ils veilleront à :

- **Définir des obligations de transparence à plusieurs niveaux tout au long de la chaîne de valeur de l'IA** (voir section 1.2).
- **Exiger une documentation pertinente et une communication claire des risques par les développeurs d'IA aux déployeurs d'IA, et par les déployeurs d'IA aux utilisateurs d'IA.** L'objectif est de garantir la sécurité de l'application et de l'utilisation des systèmes d'IA. Les déployeurs doivent ensuite imposer des garanties pour atténuer les risques susceptibles de survenir au cours de l'exploitation de ces systèmes (voir section 2).
- **Favoriser les mécanismes de prévention et de réaction rapide.** Du fait de l'ampleur et de la rapidité avec lesquelles les systèmes d'IA fonctionnent, il est indispensable de mettre en place des mécanismes de prévention et de réaction rapide aux préjudices éventuels (voir chapitre 1, section 2.3). Ce point est particulièrement important dans le contexte de la mise en place d'un système de responsabilisation en cas de violations des droits humains impliquant des technologies d'IA.[170]

168 DUDH, Articles 7, 8, and 10.
169 Par exemple, le règlement sur les services numériques définit différents régimes de responsabilité pour différents «les services de la société de l'information» («Chapitre 2 : Responsabilité des prestataires de services intermédiaires», Articles 4 à 10).
170 Yeung, K. (2019). *Responsibility and AI*, Council of Europe study. Disponible sur : https://rm.coe.int/responsability-and-ai-en/168097d9c5 (Consulté le 8 février 2024).

Dans le cadre de la définition des devoirs des développeurs et des déployeurs d'IA, les responsables politiques devraient instaurer les obligations légales suivantes :

- **Respecter les droits humains dans toutes les activités commerciales, ainsi que les principes éthiques** tels qu'énoncés par l'UNESCO dans sa *Recommandation sur l'éthique de l'intelligence artificielle*.[171]
- **Identifier et limiter les risques pour les droits humains liés à un contenu et à une utilisation préjudiciables, légaux ou illégaux,** en adoptant les meilleures pratiques, notamment par :
  - L'étiquetage humain pour l'apprentissage de l'IA (voir chapitre 1, section 1.2) ; et
  - Des activités de red-teaming utilisant des méthodologies de pointe (voir chapitre 1, section 2.1).
- **Mettre en place des mécanismes robustes pour traiter les réclamations et les réactions des utilisateurs et des autres parties concernées, et remédier aux éventuelles atteintes aux droits humains** (voir section 4.3).[172]
- **Respecter les normes de sécurité et de protection les plus récentes dans le domaine de l'IA,** en mettant continuellement à jour leurs pratiques pour qu'elles intègrent les progrès les plus récents et les meilleures pratiques dans ce domaine.
- **Assurer la transparence du système d'IA au moyen d'une approche à plusieurs niveaux.** Cela suppose de divulguer divers aspects des systèmes d'IA à différents niveaux, qu'il s'agisse d'informations techniques détaillées destinées aux autorités de réglementation et aux chercheurs agréés, ou de documents opérationnels et relatifs aux incidences accessibles et intelligibles au grand public[173] (voir chapitre 4, section 4.1).

> Par ailleurs, les responsables politiques devraient imposer aux développeurs d'IA l'obligation légale de :

- **Communiquer toutes les limitations,** y compris les restrictions d'utilisation, **aux déployeurs et aux autorités de régulation** afin d'éviter les abus, à la fois sous forme de documentation et d'accords contractuels.

> Par ailleurs, les responsables politiques devraient imposer aux déployeurs d'IA l'obligation légale de :

- **Prévoir des mesures d'atténuation des risques et de modération des résultats** dans les outils d'IA générative (voir chapitre 1, section 2.3).
- **Mentionner clairement les capacités, les limites et l'utilisation prévue des systèmes d'IA** aux utilisateurs potentiels et au public dans les conditions d'utilisation.
- **Réévaluer régulièrement les systèmes d'IA après leur mise en service** afin d'identifier et de limiter les répercussions négatives involontaires.

171 UNESCO (2021). Recommandation sur l'éthique de l'intelligence artificielle. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Consulté le 7 février 2024).

172 Les critères d'efficacité de ces mécanismes sont décrits dans l'HCDH. Access to remedy and the technology sector: basic concepts and principles, UN B-Tech Foundational Paper. Disponible sur : www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf (Consulté le 8 février 2024).

173 UNESCO (2021). Recommandation sur l'éthique de l'intelligence artificielle. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Consulté le 7 février 2024).

S'agissant de la définition des obligations des utilisateurs de systèmes d'IA dans le cadre d'activités personnelles, les responsables politiques devraient encourager :

- **Une utilisation éclairée.** Les utilisateurs devraient être encouragés à rechercher et à consulter de manière proactive des informations sur les capacités et les limites des systèmes d'IA avec lesquels ils interagissent.
- **Une utilisation responsable des systèmes d'IA conformément au droit, aux normes et aux standards internationaux en matière de droits de l'homme et aux législations nationales correspondantes.**
- **Le signalement des dysfonctionnements.** Les utilisateurs doivent être encouragés à faire part de tout dysfonctionnement ou conséquence dommageable potentiels des systèmes d'IA (voir section 4.3 et chapitre 4, section 3.1).

> **S'agissant de la définition des devoirs des utilisateurs de systèmes d'IA dans le cadre d'activités professionnelles, les responsables politiques devraient imposer l'obligation légale de :**

- **Respecter les normes éthiques et les codes de conduite professionnels** (voir chapitre 3, section 1.1).
- **Indiquer quand et dans quel but l'IA est utilisée.** Cette démarche contribuerait à préserver la transparence et la responsabilité, en particulier dans les secteurs où la confiance et la fiabilité sont primordiales.
- **Instaurer des mécanismes de recours pour les sujets de l'IA et des mécanismes de retour d'information pour signaler tout dysfonctionnement ou toute conséquence dommageable potentiels des systèmes d'IA.**

1.2. CADRES DE TRANSPARENCE MULTINIVEAUX[174][175]

Dans ses *Recommandations sur l'éthique de l'intelligence artificielle*, l'UNESCO souligne que la transparence multiniveaux (c.-à-d. la divulgation de divers aspects des systèmes d'IA à différents niveaux, des détails techniques aux informations plus générales sur leur fonctionnement et leurs effets) et la capacité à expliquer sont directement liées à la garantie de la responsabilité.[176]

En favorisant la diffusion d'informations précises sur l'ensemble du système, la transparence multiniveaux peut aider à comprendre, prédire, voire à prévenir les dommages causés par les systèmes d'IA.

Par ailleurs, cette transparence à plusieurs niveaux peut jouer un rôle fondamental en obligeant les parties responsables à rendre compte de leurs actes, et en déterminant leurs responsabilités.[177]

174 UNESCO (2023), Multilevel and Meaningful Transparency in Algorithmic Systems: Developing Concrete Criteria to Guide Institutional and Legal Reforms. Disponible sur : www.unesco.org/en/articles/multilevel-and-meaningful-transparency-algorithmic-systems-developing-concrete-criteria-guide (Consulté le 7 février 2024).

175 Belli, L. et al. (2022). Towards meaningful and interoperable transparency for digital platforms, Internet Governance Forum. Disponible sur : www.intgovforum.org/en/filedepot_download/57/23886 (Consulté le 7 février 2024).

176 UNESCO (2021). Recommandation sur l'éthique de l'intelligence artificielle. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Consulté le 7 février 2024).

177 La proposition de règlement de l'UE sur l'IA prévoit diverses obligations de transparence pour les systèmes d'IA afin de garantir que leur fonctionnement soit suffisamment clair pour les utilisateurs (article 13).

RECOMMANDATIONS
AUX **ÉTATS**

> **Imposer aux développeurs et déployeurs d'IA d'équiper les systèmes d'IA de mécanismes rigoureux d'enregistrement des informations relatives au fonctionnement de la technologie** (c.-à-d. via la « journalisation dès la conception » ou une solution technique équivalente).[178] **Ces mécanismes devraient être entièrement conformes aux lois et réglementations applicables en matière de protection des données et des secrets industriels.** Une telle exigence est essentielle pour déterminer si et quand un risque lié à la technologie se manifeste.[179]

2. RENFORCER LA RESPONSABILITÉ CONTRACTUELLE POUR LES SYSTÈMES D'IA

En l'absence de cadre réglementaire régissant les responsabilités en matière d'IA, les entreprises d'IA ont tâché de se soustraire à toute responsabilité et de réduire le degré d'incertitude via des accords contractuels. Néanmoins, les dispositions relatives à la répartition des risques dans ces contrats sont souvent biaisées en faveur des développeurs d'IA, entraînant des inégalités de pouvoir.[180]

Afin de lever ces ambiguïtés et de veiller à ce que les parties responsables soient dûment tenues de rendre des comptes conformément aux principes démocratiquement adoptés, les responsables politiques et/ou le pouvoir judiciaire devraient clarifier le cadre réglementaire régissant ces accords contractuels, ainsi que les éléments nécessaires à leur application. Cette démarche devrait également tenir compte de la nature émergente des systèmes d'IA dans divers secteurs (y compris celui des médias), de l'inadéquation des modèles de responsabilité traditionnels pour les systèmes d'IA, de la nécessité de prévoir des modes de documentation adéquats et de l'importance d'une communication transparente au sujet des risques. Ces considérations sont fondamentales pour répartir efficacement les responsabilités en cas de problème.

178 La proposition de règlement de l'UE sur l'IA oblige les entités d'IA à assurer la traçabilité des systèmes d'IA en permettant techniquement aux systèmes d'enregistrer automatiquement les événements et de conserver les journaux (article 12, article 20).
179 Union Européenne (2019), Liability for artificial intelligence and other emerging digital technologies, disponible sur : https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en (Consulté le 7 février 2024).
180 Tanenbaum,W., Song, K. and Malek, L. (2022). Theories of AI liability: It's still about the human element, Reuters. Disponible sur : www.reuters.com/legal/litigation/theories-ai-liability-its-still-about-human-element-2022-09-20/ (Consulté le 7 février 2024).

RECOMMANDATIONS AUX **ÉTATS**

> **Préciser qu'en aucun cas la responsabilité en matière d'IA ne doit être contractuellement limitée ou exclue d'une manière qui enfreint le droit, les normes et les principes internationaux en matière de droits humains, ou qui contourne les protections juridiques fondamentales.** Cette disposition vise à empêcher les développeurs d'IA de profiter des accords contractuels pour bafouer les droits fondamentaux et de se soustraire à leurs responsabilités.

> **Exiger que les accords contractuels entre les développeurs et les déployeurs d'IA précisent les droits et obligations mutuels, y compris :**

- **Les responsabilités en matière de gestion des risques.**
- **L'attribution de la responsabilité juridique.**
- **Une déclaration des déployeurs d'IA sur l'utilisation prévue d'un système d'IA avant d'en obtenir l'accès.**
- **Des dispositions permettant d'annuler l'accès si l'utilisation réelle diverge de l'objectif déclaré.**
- **Les limites du système et les lacunes potentielles.**

> **Exiger que les conditions d'utilisation des systèmes d'IA précisent :**

- **La responsabilité des résultats et des décisions de l'IA.** En règle générale, la responsabilité des résultats et des décisions d'un système d'IA est imputable à son déployeur et à son développeur s'ils ne respectent pas les mesures requises en matière d'atténuation des risques, de transparence, de diligence et pour les mécanismes de notification et d'action (comme indiqué à la section 3.1).[181] Ils sont également responsables de la rectification des résultats irréguliers et préjudiciables. Les utilisateurs, en revanche, devraient être tenus pour responsables, conformément à la législation en vigueur, de tout préjudice causé intentionnellement par l'utilisation de systèmes d'IA.
- **La propriété des données et du contenu.** Les utilisateurs restent propriétaires des données et des contenus partagés avec un système d'IA. De plus, ils peuvent refuser que leurs données et les métadonnées de leurs interactions soient stockées et réutilisées pour optimiser le système. Cette clarification ne doit pas être utilisée par les déployeurs d'IA pour contourner leur obligation de se conformer aux réglementations en vigueur en matière de droits d'auteur et de protection des données. Au contraire, ils devraient être tenus d'indiquer explicitement dans quelle mesure leurs systèmes d'IA respectent ces lois et de garantir un traitement transparent des données entrantes et des données sortantes (voir chapitre 1, section 1.1.e.).
- **Les cas d'utilisation interdits pour l'application des systèmes d'IA, conformément au droit international des droits humains.**[182]
- **Les limites du système et les lacunes potentielles**.

181 Le règlement de l'UE sur les services numériques oblige les fournisseurs de services d'hébergement à permettre aux utilisateurs de leur notifier la présence d'informations qu'ils considèrent comme illicite, puis de traiter ces notifications (article 16).

182 Voir les pratiques interdites en matière d'IA dans la proposition de règlement de l'UE sur l'IA (article 5.1). En outre, selon le règlement de l'UE sur les services numériques, les plateformes en ligne doivent suspendre leurs services aux utilisateurs qui fournissent du contenu illicite (article 23).

- **La disponibilité et le fonctionnement des mécanismes de retour d'information des utilisateurs** (voir chapitre 1, section 2.3, et chapitre 4, section 3.1) **et des mécanismes de recours** (voir section 4.3).[183]

3. IDENTIFIER DES RÉGIMES DE RESPONSABILITÉ APPROPRIÉS POUR LES SYSTÈMES D'IA

L'opacité et la complexité des systèmes d'IA sont telles qu'elles peuvent entraîner une attribution de la responsabilité inéquitable, inefficace, voire impossible, ou empêcher un sujet d'IA victime d'un préjudice ou d'un dommage d'en prouver la causalité et le priver d'indemnisation ou de recours.[184]

Afin de résoudre ces problèmes, les responsables politiques devraient réévaluer les lois existantes en matière de responsabilité et, le cas échéant, adopter de nouvelles réglementations afin de clarifier leur applicabilité aux systèmes d'IA.

Tout d'abord, il convient de préciser que dans le cas d'actions en justice intentées à l'encontre de développeurs et de déployeurs d'IA par des individus ou un groupe ayant subi des préjudices, la charge de la preuve devrait revenir aux développeurs et aux déployeurs d'IA.

En outre, en vue de déterminer le régime de responsabilité le plus approprié pour les systèmes d'IA dans l'espace de l'information et de la communication, les responsables politiques devraient prendre plusieurs facteurs en compte. Ceux-ci comprennent les fonctions des systèmes d'IA, la nature, la gravité, la probabilité et la réversibilité du préjudice causé, ainsi que le devoir de diligence des développeurs et des déployeurs de l'IA (voir le tableau 2.1). À l'inverse, les initiatives actuelles visant à différencier les systèmes d'IA prédictifs et génératifs, quoique bien intentionnées, risquent de simplifier à l'excès le monde complexe de la technologie de l'IA, la frontière entre ces deux types de systèmes s'avérant de plus en plus floue.

183 Le règlement de l'UE sur les services numériques oblige les plateformes en ligne à fournir aux utilisateurs un système interne de traitement des réclamations (article 20).

184 Parlement Européen (2020), Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission sur un régime de responsabilité civile pour l'intelligence artificielle. Disponible sur : https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_FR.html (Consulté le 8 février 2024).

Tableau 2.1. Classification des systèmes d'IA utilisés dans l'espace de l'information et de la communication, des préjudices potentiels et des régimes de responsabilité

FONCTIONS	EXEMPLES	PRÉJUDICES POTENTIELS	CONSIDÉRATIONS SUR LA RESPONSABILITÉ
Vérification et modération des contenus (traditionnellement effectuées par des systèmes d'IA prédictive)	• Vérification des faits • Détection de spam • Analyse d'images/ vidéos • Détection des propos haineux • Signalement de contenu	• Préjugés • Censure • Menaces négligées • Violations de la vie privée • Absence de procédure adéquate	Les décisions prises par les algorithmes de vérification et de modération des contenus peuvent entraîner des violations des droits humains. En conséquence, les développeurs et déployeurs d'IA devraient être tenus responsables des préjudices qu'ils causent s'ils ne respectent pas les mesures d'atténuation des risques, les exigences de transparence et le devoir de diligence, sauf preuve du contraire. La responsabilité devrait également s'étendre aux cas où ils ne parviennent pas à établir et à gérer efficacement un mécanisme clair et prévisible de recours contre les mesures de modération des contenus.
Curation de contenu et recommandation (traditionnellement réalisées par des systèmes d'IA prédictive)	• Moteurs de recommandation • Agrégateurs d'actualités personnalisées • Algorithmes de référencement des résultats de recherche	• Bulles de filtre et chambres d'écho • Polarisation • Manipulation	Compte tenu de la nature subjective des décisions prises par les algorithmes de curation et de recommandation, les résultats préjudiciables n'indiquent pas nécessairement un dysfonctionnement du système. Par conséquent, les développeurs et déployeurs d'IA devraient être tenus pour responsables du non-respect des mesures d'atténuation des risques, des exigences de transparence et du devoir de diligence, sauf preuve du contraire.
Ciblage et diffusion des annonces (traditionnellement effectués par des systèmes d'IA prédictifs)	• Microciblage • Ciblage contextuel	• Manipulation • Discrimination • Violation de la vie privée • Tromperie	Étant donné que les préjudices potentiels résultent principalement du microciblage fondé sur des caractéristiques protégées et des catégories particulières de données à caractère personne[185]l, les développeurs d'IA et les déployeurs d'algorithmes utilisés pour le microciblage fondé sur la race, la couleur, le sexe, l'orientation sexuelle, l'identité de genre, la langue, la religion, les opinions politiques ou autres, l'origine nationale ou sociale, le patrimoine, la naissance ou tout autre statut protégé, ainsi que les catégories particulières de données à caractère personnel considérées comme nécessitant une protection plus élevée, devraient être tenus responsables de tout préjudice ou dommage qui en résulterait. Cette responsabilité devrait s'appliquer indépendamment du fait qu'ils aient eu l'intention de causer un préjudice ou qu'ils aient agi par négligence.[186] À l'inverse, les développeurs d'IA et les déployeurs d'algorithmes utilisés pour le ciblage comportemental ou contextuel devraient être tenus pour responsables du non-respect des mesures d'atténuation des risques, des exigences de transparence et du devoir de diligence requis, sauf preuve du contraire.
Création de contenu (systèmes d'IA générative)	• Outils rédactionnels d'IA • Générateurs d'images • Générateurs d'audio • Générateurs de vidéo • Bots de réseaux sociaux pilotés par l'IA	• Mésinformation et désinformation / deepfakes • Violations des droits d'auteur • Violations de la vie privée et de la protection des données • Discours de haine et harcèlement • Propos illégaux • Usurpation d'identité	Bien que les développeurs et les déployeurs ne soient pas en capacité d'éliminer complètement le risque d'utilisation abusive des systèmes d'IA générative, la mise en œuvre de mesures de protection peut considérablement réduire le risque de violations des droits humains. Afin d'encourager le respect des pratiques de sécurité et d'éthique, les développeurs et les déployeurs devraient être tenus pour responsables en cas de non-respect des mesures d'atténuation des risques, des exigences de transparence et du devoir de diligence, sauf preuve du contraire. Par ailleurs, les développeurs et déployeurs d'IA devraient être tenus pour strictement responsables s'ils ne donnent pas suite aux notifications et ne rectifient pas les systèmes en conséquence et en temps utile concernant les résultats illégaux (protection des données, droits d'auteur, diffamation, discours haineux, incitation à la violence, etc.). Il faut poursuivre les discussions avec la société civile, les universitaires et les défenseurs des droits humains pour déterminer si un régime de responsabilité stricte devrait être applicable dans les cas de violation des droits fondamentaux. Ces discussions doivent également prendre en compte les contextes juridiques et culturels uniques des différentes juridictions.
Personnalisation (systèmes d'IA générative)	• Chatbots • Assistants virtuels	• Manipulation • Surveillance/ violation de la vie privée	

185 Les catégories spéciales de données à caractère personnel identifiées par la législation nationale ou régionale comme nécessitant une protection renforcée comprennent les données révélant l'opinion politique, l'appartenance syndicale, les données relatives à la santé, les données génétiques et biométriques (voir par exemple le RGPD européen).

186 Des discussions supplémentaires sont nécessaires pour définir des mesures législatives appropriées concernant la publicité basée sur la surveillance et reposant sur le suivi comportemental, comme discuté dans l'OSCE (2021). Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual. Disponible sur : www.osce.org/files/f/documents/8/f/510332_1.pdf (Consulté le 8 février 2024).

Enfin, compte tenu du rôle important joué par les plateformes qui hébergent des contenus générés par l'IA (telles que les moteurs de recherche et les réseaux sociaux) dans la diffusion et la visibilité de contenus potentiellement préjudiciables, il est urgent de définir des mécanismes de responsabilité spécifiques et des mesures d'atténuation des risques.[187]

RECOMMANDATIONS AUX **ÉTATS**

> **Réviser les lois applicables en matière de responsabilité ou adopter des régimes spéciaux de responsabilité afin de clarifier la manière dont elles s'appliquent aux systèmes d'IA. Dans ce cadre, les responsables politiques devraient :**
>
> - **Préciser que toutes les entités impliquées sur toute la chaîne de valeur de l'IA, y compris celles qui créent, entretiennent ou contrôlent les risques associés au système d'IA, sont en principe responsables des résultats et des décisions de l'IA.** Ce principe répond à la difficulté d'identifier les résultats ou décisions préjudiciables de l'IA par rapport à des intrants humains ou des choix de conception spécifiques, et de veiller à ce que les victimes reçoivent une indemnisation ou une réparation.[188]
> - **Préciser que la charge de la preuve devrait échoir aux développeurs et déployeurs d'IA en cas d'actions en justice intentées à leur encontre lorsque des individus ou un groupe ont subi un préjudice.**
> - **Compte tenu de la subjectivité inhérente à la définition du « préjudice », impliquer activement des organisations de la société civile, des universitaires et des chercheurs de diverses disciplines, origines et pays dans le processus de définition de ce concept.**
> - **Établir une distinction entre la fonction prévue du système d'IA (vérification et modération du contenu, curation de contenu et recommandation, ciblage et diffusion de publicités, création de contenu, personnalisation), la nature, la gravité, la probabilité et la réversibilité du préjudice causé, et le devoir de diligence des développeurs et des déployeurs d'IA dans les cadres juridiques.**

3.1. DÉVELOPPEURS ET DÉPLOYEURS DE SYSTÈMES D'IA DESTINÉS À UNE UTILISATION DANS L'ESPACE INFORMATIONNEL

Les systèmes d'IA assument un rôle décisionnel de plus en plus important en termes de modération, de vérification, de curation et de recommandation d'informations, mais aussi en matière de ciblage et de diffusion d'annonces publicitaires. Ces applications varient suivant leur fonction et leur utilisation spécifique, ce qui engendre des risques différents. La gravité de ces risques devrait être directement associée au type de responsabilité imposée aux développeurs et aux utilisateurs de l'IA, comme indiqué dans le tableau 2.1.

187 Bien que le règlement de l'UE sur les services numériques prévoie certains éléments en ce sens, il manque de dispositions spécifiques à l'IA générative, dans la mesure où il se concentre sur la gouvernance des plateformes en ligne et des moteurs de recherche.

188 Parlement Européen (2020), Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission sur un régime de responsabilité civile pour l'intelligence artificielle. Disponible sur : https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_FR.html (Consulté le 8 février 2024). Cela soulignerait également qu'il existe une responsabilité solidaire parmi les différents acteurs de la chaîne de valeur de l'IA.

Par exemple, le microciblage des utilisateurs en fonction de la race, de la couleur, du sexe, de l'orientation sexuelle, de l'identité de genre, de la langue, de la religion, des opinions politiques ou autres, de l'origine nationale ou sociale, de la propriété, de la naissance ou de tout autre statut présente des risques importants pour les droits fondamentaux. Ces risques englobent des problématiques telles que la manipulation, la discrimination, la violation de la vie privée et la tromperie. Au regard de ces préoccupations, les développeurs et déployeurs de systèmes d'IA utilisés pour le ciblage et la diffusion de publicités fondées sur des critères protégés et des catégories particulières de données à caractère personnel devraient être tenus responsables de tout préjudice ou dommage causé par ces systèmes, indépendamment de la présence ou de l'absence de négligence ou d'acte répréhensible intentionnel. À l'inverse, en raison de l'atteinte moins importante à la vie privée que représentent les systèmes d'IA utilisés pour le ciblage comportemental ou contextuel, leurs développeurs et déployeurs devraient être tenus pour responsables du non-respect des mesures d'atténuation des risques, des exigences de transparence et du devoir de diligence requis, sauf preuve du contraire, ou s'il peut être démontré qu'ils ont commis des actes répréhensibles, portant ainsi préjudice ou dommage à une tierce partie.

Parallèlement, il convient de reconnaître que dans certains cas, malgré le risque de violation des droits humains, les développeurs et les utilisateurs de l'IA sont confrontés à des choix délicats, et doivent choisir entre la minimisation du risque de faux positifs ou de faux négatifs. C'est pourquoi les développeurs et les déployeurs de systèmes d'IA utilisés pour la vérification et la modération des contenus ne devraient pas être soumis à une responsabilité stricte. Toutefois, ils devraient être tenus pour responsables du non-respect des mesures d'atténuation des risques, des exigences de transparence et du devoir de diligence, sauf preuve du contraire, ainsi que de l'absence de mise en place et de gestion efficace d'un dispositif clair et prévisible de recours contre les mesures de modération des contenus.[189]

Ces considérations devraient s'étendre aux développeurs et aux déployeurs de systèmes d'IA utilisés pour la curation et la recommandation de contenu. Dans de tels cas, la nature subjective des décisions prises par ces algorithmes ne reflète pas nécessairement un dysfonctionnement du système. Les développeurs et déployeurs devraient donc être tenus responsables du non-respect des mesures d'atténuation des risques, des exigences de transparence et du devoir de diligence, sauf preuve du contraire, ou s'il peut être prouvé qu'ils ont commis des actes répréhensibles ayant porté préjudice ou causé des dommages à des utilisateurs.

Les systèmes d'IA utilisés pour la création et la personnalisation de contenu présentent des difficultés spécifiques et supplémentaires au regard des régimes de responsabilité existants. Ces difficultés incluent notamment la prise en compte du fait que même des mesures de protection judicieusement conçues peuvent être contournées, la répartition des responsabilités lorsque les mesures de protection échouent et que les systèmes d'IA générative sont exploités à des fins préjudiciables,[190] et l'établissement d'une norme de diligence raisonnable pour les développeurs et les déployeurs de tels systèmes.

Pour favoriser le respect des pratiques sécuritaires et éthiques, les responsables politiques pourraient exiger que les développeurs et les déployeurs de systèmes d'IA utilisés pour la création et la personnalisation de contenus soient tenus responsables en cas de non-respect de leurs obligations liées aux mesures d'atténuation des risques (y compris la modération des résultats), aux exigences de transparence et au devoir de diligence, sauf preuve du contraire. De plus, les développeurs et les déployeurs de systèmes d'IA utilisés pour la création et la personnalisation de contenus devraient être tenus responsables en cas de gestion inefficace des procédures de notification et de retrait des résultats

189 Pour plus d'information : Forum sur l'information et la démocratie (2022). Régimes de Responsabilité pour les Réseaux Sociaux et leurs Utilisateurs. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Responsabilite-reseaux-sociaux_FR.pdf (Consulté le 8 février 2024).

190 Henderson, P. (2023). Who Is Liable When Generative AI Says Something Harmful? Stanford University. Disponible sur : https://hai.stanford.edu/news/who-liable-when-generative-ai-says-something-harmful (Consulté le 8 février 2024).

illégaux, ce qui comprend des questions telles que la protection des données, les violations des droits d'auteur, la diffamation, les discours de haine, l'incitation à la violence, etc.

Des discussions plus approfondies avec la société civile, la communauté universitaire et les défenseurs des droits humains sont nécessaires pour déterminer si un régime de responsabilité stricte devrait également être applicable dans les cas de violation des droits fondamentaux. Il est essentiel que ces échanges prennent en compte les contextes juridiques et culturels uniques propres à chaque juridiction.

Enfin, au sein des juridictions où les plateformes ne sont généralement pas responsables du contenu des tiers, les responsables politiques devraient spécifier que les systèmes d'IA générant du contenu ne sont pas automatiquement couverts par cette immunité générale.

RECOMMANDATIONS AUX **ÉTATS**

> **Instaurer un régime de responsabilité stricte pour les développeurs et les déployeurs de systèmes d'IA visant à microcibler les utilisateurs sur la base de critères protégés** tels que la race, la couleur, le sexe, l'orientation sexuelle, l'identité de genre, la langue, la religion, les opinions politiques ou autres, l'origine nationale ou sociale, la propriété ou la naissance, **ainsi que des catégories particulières de données à caractère personnel requérant une protection renforcée** (p. ex. les données divulguant l'opinion politique, l'appartenance à un syndicat, les données relatives à la santé, les données génétiques et les données biométriques). Dans ce cadre, les développeurs et déployeurs d'IA devraient être tenus responsables de tout préjudice ou dommage causé par ces systèmes, indépendamment de leur culpabilité. **Le microciblage fondé sur des critères protégés et des catégories particulières de données à caractère personnel devrait être interdit.**[191]

> **Instaurer un régime de responsabilité fondé sur la faute pour les développeurs et les déployeurs de systèmes d'IA destinés à cibler et à diffuser des publicités fondées sur le comportement en ligne (c.-à-d. la publicité comportementale) ou sur le contexte (c.-à-d. la publicité contextuelle).** Dans ce cadre, les développeurs et les déployeurs de systèmes d'IA devraient être tenus responsables du non-respect de leurs obligations liées aux mesures d'atténuation des risques, aux exigences de transparence et au devoir de diligence, sauf preuve du contraire.

> **Instaurer un régime de responsabilité fondé sur la faute pour les développeurs et les déployeurs de systèmes d'IA utilisés pour la vérification, la modération et la recommandation de contenus.** Dans ce cadre, les développeurs et les déployeurs de systèmes d'IA ne seraient pas automatiquement tenus responsables des préjudices causés par leurs systèmes. Toutefois, ils seraient tenus pour responsables du non-respect des obligations liées aux mesures d'atténuation des risques, aux exigences de transparence et au devoir de diligence, sauf preuve du contraire. En outre, ils seraient également responsables en cas d'absence de mise en place ou de fonctionnement

191 Le règlement de l'UE sur les services numériques réglemente la publicité microciblée de trois manières : premièrement, il interdit le microciblage (« profilage ») basé sur les catégories particulières de données à caractère personnel (article 26.3). Deuxièmement, elle interdit totalement la publicité microciblée lorsque les destinataires sont mineurs (article 27). Troisièmement, il oblige les annonceurs en ligne à fournir une explication des principaux paramètres utilisés dans leurs systèmes de recommandation, ainsi que des options permettant aux destinataires de modifier ou d'influencer ces principaux paramètres (article 27). Bien que cela sorte du cadre de ce rapport, il est recommandé aux États d'envisager une interdiction totale de la publicité microciblée destinée aux mineurs en raison de son impact significatif sur la formation de leur liberté d'expression et d'opinion.

d'un mécanisme transparent et prévisible de recours contre les actions de modération des contenus.

> **Instaurer un régime de responsabilité fondé sur la faute pour les développeurs et les déployeurs de systèmes d'IA utilisés pour la génération de résultats, tels que le contenu ou la personnalisation.** Dans ce cadre, les développeurs et déployeurs de systèmes d'IA seraient tenus responsables du non-respect de leurs obligations liées aux mesures d'atténuation des risques, aux exigences de transparence et au devoir de diligence, sauf preuve du contraire. En outre, le présupposé initial devrait être que tout préjudice causé par des systèmes d'IA générative est dû à la négligence, à moins que les développeurs et les déployeurs ne puissent prouver qu'ils se conforment à toutes les normes juridiques et réglementaires en vigueur. Ce principe garantit un plus haut degré de responsabilité et encourage le respect rigoureux des pratiques de sécurité et d'éthique en matière de développement et de déploiement de l'IA.

> **Établir un régime de responsabilité stricte imposant aux développeurs et déployeurs de systèmes d'IA utilisés pour la génération de résultats, tels que le contenu ou la personnalisation, de mettre en place et d'appliquer des procédures de notification et d'action pour les résultats de leurs systèmes.** Dans ce cadre, les développeurs et déployeurs de systèmes d'IA sont tenus responsables de tout préjudice causé par leurs systèmes s'ils ne répondent pas rapidement aux notifications et ne rectifient pas les résultats du système illicites (y compris les questions liées à la protection des données, aux droits d'auteur, à la diffamation, aux discours haineux, à l'incitation à la violence, etc.), sans exemptions.

> **Entamer un dialogue inclusif avec les différentes parties prenantes (y compris, à minima, des représentants de la société civile et des organismes de défense des droits humains, des groupes vulnérables et des journalistes) afin de déterminer dans quelles situations de violation des droits fondamentaux un régime de responsabilité stricte devrait s'appliquer aux développeurs et aux déployeurs de systèmes d'IA utilisés pour générer des résultats, tels que du contenu ou de la personnalisation.[192] Ces débats doivent également prendre en compte les contextes juridiques et culturels spécifiques à chaque juridiction.**

> **Clarify by legal means that content generated by AI should not be considered as third-party content or hosting content by the AI system in determining the liability of the generative AI deployer.**

> **Préciser par des moyens juridiques que le contenu généré par l'IA ne doit pas être considéré comme un contenu de tiers ou un contenu d'hébergement par le système d'IA pour déterminer la responsabilité de l'auteur du déploiement de l'IA générative.**

192 Cette responsabilité pourrait également s'appliquer à d'autres cas tels que la pédopornographie, mais ceci sort du cadre du présent rapport.

3.2. PLATEFORMES HÉBERGEANT DES CONTENUS ET DES ENTITÉS GÉNÉRÉS PAR L'IA

L'un des défis majeurs des plateformes (plateformes numériques ou moteurs de recherche) qui hébergent des contenus et des entités générés par l'IA consiste à atteindre un équilibre entre la préservation de la liberté d'expression et la réduction des préjudices causés par les contenus trompeurs, qu'ils soient générés par l'homme ou par l'IA.

Pour prévenir l'influence néfaste des plateformes sur l'espace public numérique et favoriser une modération et une curation responsables des contenus, il est indispensable de définir et d'établir des responsabilités claires pour les activités et les contenus préjudiciables qu'elles hébergent. Cela concerne notamment les activités préjudiciables que l'IA permet désormais, telles que la manipulation électorale, le micro-ciblage et la surveillance de masse.

Cette approche devrait s'appuyer sur la présomption de responsabilité des plateformes en cas de négligence des mesures d'atténuation des risques, des exigences de transparence, du devoir de diligence et de l'absence de mise en œuvre de normes de pointe en matière de détection, d'étiquetage, d'authenticité et de provenance.

RECOMMANDATIONS AUX **ÉTATS**

> **Instaurer une présomption réfutable en vertu de laquelle les plateformes sont responsables des contenus illégaux qu'elles hébergent (qu'ils soient générés par des êtres humains ou par l'IA)**[193] **et des préjudices qu'elles causent, à moins qu'elles ne puissent prouver qu'elles ont :**
>
> ♦ **Mis en place des mesures exhaustives d'atténuation des risques**, y compris :
>
> • Des mesures accessibles à tous les utilisateurs leur permettant de signaler un contenu ou un comportement illégal. Celles-ci comprennent l'examen et la notification en temps utile des utilisateurs concernés et de la personne qui a signalé le contenu, ainsi qu'un mécanisme de recours.[194]
>
> • Des dispositifs pour les signaleurs de confiance[195] et la coopération avec des fact-checkers indépendants.[196][197]
>
> • Des évaluations des risques systémiques, y compris les risques systémiques découlant des contenus générés par l'IA.[198]
>
> ♦ **Respecté les exigences en matière de transparence** (voir chapitre 4, section 4).

193 Cette proposition doit être rapportée au cadre juridique existant aux États-Unis, en particulier en ce qui concerne la section 230 de la loi sur la décence des communications (Communications Decency Act). Cette section a historiquement fourni une large immunité aux plateformes en ligne contre la responsabilité civile pour le contenu publié par des tiers. En dépit des nombreux problèmes rencontrés et des demandes de réforme, elle reste une pierre angulaire du droit de l'internet aux États-Unis.

194 Pour plus d'information, consultez Forum sur l'information et la démocratie (2022). Régimes de Responsabilité pour les Réseaux Sociaux et leurs Utilisateurs. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Responsabilite-reseaux-sociaux_FR.pdf (Consulté le 8 février 2024).

195 Forum sur l'information et la démocratie (2020). Pour mettre fin aux Infodémies. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Infodemies_FR.pdf (Consulté le 8 février 2024), pp.27-28 and page 37.

196 *Ibid*, pp. 60, 80-81, 100-101.

197 Le règlement de l'UE sur les services numériques oblige les plateformes en ligne à prendre les mesures nécessaires pour s'assurer que les signaleurs de confiance peuvent fonctionner de manière efficace (Article 22).

198 Pour une définition des risques systémiques, voir le chapitre 1, section 2.2. Le règlement de l'UE sur les services numériques oblige les plateformes en ligne à effectuer une évaluation des risques en mettant l'accent sur les droits humains (article 34).

♦ **Adopté les dernières normes en matière de détection, d'étiquetage, de provenance et d'authenticité** :

- Cela inclut le respect des normes existantes en matière de conservation des métadonnées ou des signatures cryptographiques permettant d'identifier la provenance et l'authenticité, ou les filigranes lors du téléchargement et du partage de contenu sur leurs plateformes.
- La détection des filigranes et des informations relatives à la provenance et à l'authenticité et la diffusion de ces informations aux utilisateurs finaux.

4. AUTONOMISER LES UTILISATEURS ET LES SUJETS DE L'IA : RESPONSABILISER EN MATIÈRE DE CRÉATION ET DE MODÉRATION DE CONTENUS SYNTHÉTIQUES

S'il est primordial d'établir la responsabilité d'une sortie ou d'une décision générée par l'IA, cela reste insuffisant pour garantir une véritable obligation de rendre des comptes. Les systèmes d'IA, en particulier ceux qui ont une incidence importante sur les individus ou la société, nécessitent davantage qu'une simple définition des tâches ou attribution des responsabilités.

Pour garantir une véritable responsabilisation, les utilisateurs et toute personne concernée par les systèmes d'IA doivent savoir comment ces systèmes fonctionnent. De plus, ils doivent être informés du moment et de la manière dont l'IA est utilisée. Enfin, ils doivent pouvoir bénéficier de droits juridiquement exécutoires en cas de conséquences négatives de l'IA.

Les recommandations formulées dans la présente section concernent essentiellement les contenus synthétiques. Nous invitons les personnes intéressées à consulter le rapport du Forum sur l'information et la démocratie *Des régimes de responsabilité pour les réseaux sociaux et leurs utilisateurs* (2022) pour une analyse détaillée des stratégies visant à responsabiliser les utilisateurs affectés par les décisions prises par les systèmes d'IA dans l'espace informationnel, notamment en ce qui concerne les décisions de modération des contenus et des comptes sur les plateformes.[199]

199 Forum sur l'information et la démocratie (2022). Régimes de Responsabilité pour les Réseaux Sociaux et leurs Utilisateurs. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Responsabilite-reseaux-sociaux_FR.pdf (Consulté le 8 février 2024).

RECOMMANDATIONS
AUX **ÉTATS**

> **Définir un cadre juridique complet définissant clairement les droits des individus dans le contexte des décisions et des résultats de l'IA. Ce cadre devrait comprendre des dispositions spécifiques pour chaque type de préjudice imputable à l'IA, et devrait inclure :**
>
> - **Le droit des personnes à être informées lorsqu'un système d'IA a été utilisé pour prendre des décisions ayant des conséquences pour elles ou pour générer des résultats les concernant.** Cela implique une explication claire du rôle et de la raison d'être de l'implication de l'IA dans ces décisions.
> - **Le droit de recevoir des explications sur les décisions et les résultats de l'IA qui soient techniquement exactes, mais présentées de manière compréhensible et pertinente pour l'utilisateur.** Ces explications devraient comprendre des informations sur les données et les critères utilisés par l'IA pour prendre ces décisions.
> - **Le droit de contester les décisions et les résultats émanant de l'IA, avec la garantie de bénéficier d'un examen humain rapide.**[200] Ce processus doit être facilement accessible et permettre de résoudre les problèmes dans les meilleurs délais.
> - **Le droit à la non-discrimination en ce qui concerne les décisions et les résultats de l'IA, garantissant que les systèmes d'IA ne reproduisent pas de biais ou de traitement inégal fondé sur la race, la couleur, le sexe, l'orientation sexuelle, l'identité de genre, la langue, la religion, l'opinion politique ou autre, l'origine nationale ou sociale, la propriété, la naissance ou d'autres caractéristiques protégées en vertu du droit international des droits humains.**

4.1. CONTENU SYNTHÉTIQUE ET ENTITÉS SYNTHÉTIQUES

L'intégration de l'IA dans la production de textes, de vidéos, d'images et de contenus audio suscite d'importantes préoccupations en matière de transparence et d'authentification des contenus dans l'espace informationnel. Bien souvent, les utilisateurs interagissent avec des contenus et des entités générés par l'IA sans en connaître clairement l'origine, ni connaître la personne à l'origine du système, la configuration technique et le fonctionnement du système (p.ex. les jeux de données, les indicateurs d'apprentissage).

Afin de garantir la confiance et la responsabilité, les utilisateurs et les sujets de l'IA doivent être informés du moment et de la manière dont l'IA est utilisée. Il faut donc favoriser la transparence sur les biais potentiels, les limites et la précision des systèmes d'IA avec lesquels ils interagissent.

200 La proposition de règlement de l'UE sur l'IA exige un contrôle humain des systèmes d'IA à haut risque (article 14).

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux systèmes d'IA d'intégrer un filigrane indiquant la provenance dans tous les contenus synthétiques qu'ils génèrent, y compris les textes, les vidéos, les images et les sons, et de prévoir des méthodes permettant de détecter de manière fiable ce type de contenu** (voir chapitre 1, section 1.5)[201]. **En outre, il faut exiger que tous les distributeurs successifs de ces contenus conservent le filigrane afin de garantir la transparence de leur origine.**

> **Imposer aux déployeurs d'entités synthétiques (p.ex. chatbots, assistants virtuels) d'informer les utilisateurs qu'ils interagissent avec un système interactif piloté par l'IA.**[202]

> **Imposer aux plateformes distribuant des contenus synthétiques et hébergeant des entités synthétiques d'adopter une politique exhaustive en matière d'utilisation de contenus et de comptes synthétiques, et de rendre cette politique facilement accessible et clairement compréhensible.** Cette politique devrait inclure :

- Des exigences relatives à l'étiquetage des contenus générés par l'IA, en distinguant les contenus (photo)réalistes ou d'apparence authentique, les contenus prêtant à confusion ou trompeurs, les contenus générés par l'IA utilisés à des fins satiriques et artistiques, les contenus générés par l'IA par les médias et les contenus générés par l'IA dans les annonces politiques et au sujet de personnalités très influentes.
- L'identification des types de contenus qui sont interdits et considérés comme illégaux et de ceux qui feront l'objet d'un avertissement.
- Des dispositifs de signalement facilement accessibles pour pointer la nature synthétique d'un contenu ou d'un compte.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Informer les utilisateurs qu'ils interagissent avec des entités synthétiques, telles que les chatbots ou les assistants virtuels, en étiquetant clairement l'interface utilisateur de l'IA ou en présentant un avertissement avant chaque interaction.**

> **En outre, il convient d'ajouter à ces mécanismes de transparence un avertissement explicite concernant les inexactitudes, les biais ou les contre-vérités potentielles du contenu généré par des entités synthétiques (c.-à-d. les « hallucinations de l'IA »). Cette mesure permettra aux utilisateurs de procéder à une évaluation critique du contenu généré par l'IA.**

201 La proposition de règlement de l'UE sur l'IA oblige les fournisseurs de systèmes d'IA, y compris les systèmes d'IA à usage général, à concevoir des systèmes de telle façon que les images, textes, contenus audio et vidéo à caractère synthétique fassent l'objet d'un marquage dans un format lisible par machine et que leur nature de contenus générés ou manipulés artificiellement soit détectable (Article 52).

202 De la même manière, la proposition de règlement de l'UE sur l'IA prévoit que les systèmes d'IA destinés à interagir directement avec les humains devraient être conçus de manière à informer les humains qu'ils interagissent avec un système d'IA (article 52.1). En outre, les fournisseurs de systèmes d'IA qui génèrent du contenu synthétique doivent s'assurer que celui-ci est spécifiquement désigné comme étant généré artificiellement (article 57.2).

RECOMMANDATIONS AUX **PLATEFORMES**

> **Adopter une politique exhaustive sur l'utilisation de contenus et de comptes synthétiques, et veiller à ce que cette politique soit facilement et clairement compréhensible pour les utilisateurs. Elle devrait être élaborée dans le cadre d'un processus participatif et inclusif prévoyant à minima une participation équitable, durable et substantielle de chercheurs indépendants, de la société civile et de groupes marginalisés. Ce processus et la sélection des parties prenantes doivent impérativement être transparents.**

> **Mettre en œuvre une politique en vertu de laquelle le contenu et les utilisateurs doivent se voir accorder un « droit de recommandation » avant de faire l'objet d'une promotion ou de figurer dans les fils d'actualité. Ce droit devrait être accordé sur la base d'une signature cryptographique valide associée à des entités de confiance.** Une telle politique répond à la nécessité de sélectionner avec soin les objectifs d'optimisation dans le développement des systèmes de recommandation afin de préserver l'intégrité de l'écosystème de l'information numérique, question qui est examinée plus en détail au chapitre 1, section 1.4.

> **Accroître la part de contenu authentifié dans leurs systèmes de recommandation,** en particulier les médias d'intérêt public certifiés par la Journalism Trust Initiative.[203]

4.2. UTILISATION DE L'IA GÉNÉRATIVE EN POLITIQUE ET AUTRES CONTEXTES HAUTEMENT SENSIBLES

Les chatbots politiques et les contenus politiques générés par l'IA, ainsi que l'utilisation plus étendue de l'IA générative dans des contextes très sensibles (p.ex. pandémies, élections, périodes de conflit), devraient être soumis aux mêmes exigences de transparence et aux mêmes limites que les contenus créés par l'humain.

Dans cette optique, les États doivent réévaluer et mettre à jour les lois existantes (notamment les lois électorales) chaque fois que cela est nécessaire pour faire face aux difficultés spécifiques que pose l'IA dans le cadre des campagnes politiques, et instaurer une réglementation appropriée.[204] Il s'agit notamment de réglementer les deepfakes et les chatbots utilisés par les partis politiques et les gouvernements pour communiquer avec les électeurs ou diffuser des informations sur les processus électoraux ou les questions politiques/de campagne.

En parallèle, les plateformes devraient réévaluer en permanence leurs outils de lutte contre la tromperie numérique, afin de s'assurer de leur conformité avec la législation. Cette réévaluation devrait porter sur l'efficacité des politiques et des solutions techniques mises en place pour faire face aux menaces

203 Journalism Trust Initiative. Disponible sur : www.journalismtrustinitiative.org/ (Consulté le 8 février 2024).

204 Les lignes directrices provisoires du règlement de l'UE sur les services numériques sur l'intégrité des élections élaborées par la Commission européenne constituent une avancée dans cette direction, Commission européenne (2024). Commission is gathering views on draft DSA guidelines for election integrity. Disponible sur : https://digital-strategy.ec.europa.eu/en/news/commission-gathering-views-draft-dsa-guidelines-election-integrity (Consulté le 15 février 2024).

émergentes que les technologies d'IA avancées font peser sur la liberté d'expression et d'opinion, ainsi que sur l'intégrité des débats politiques.

RECOMMANDATIONS AUX **ÉTATS**

> **Mettre à jour les lois électorales afin de définir des règles claires sur l'utilisation de l'IA générative et des deepfakes dans les campagnes électorales. Ces réglementations devraient spécifiquement imposer des exigences de divulgation et de transparence concernant l'utilisation de ces technologies.**

> **Imposer aux plateformes de se doter de techniques complètes d'atténuation des risques, en particulier pour les contenus générés par l'IA en rapport avec les élections et d'autres contextes très sensibles** [205](voir chapitre 1, section 2.3).

> **Imposer des évaluations de red-teaming et des risques sur mesure, et instaurer des mécanismes de vérification renforcés pour les systèmes d'IA (tels que les chatbots) devant être utilisés dans des contextes politiques et d'autres circonstances très sensibles. Ces mesures tiennent compte des effets préjudiciables que les hallucinations ou les résultats polarisants, biaisés et discriminatoires peuvent avoir dans de tels contextes** (voir chapitre 1, sections 2.1 et 2.2).

> **Dans le cadre de la réglementation de l'utilisation de l'IA lors d'élections et de processus politiques, il est essentiel de privilégier la manière dont le contenu est transmis au public, plutôt que le contenu lui-même. À cet égard, l'interdiction du micro-ciblage fait partie des mesures efficaces** [206](voir section 3.1).

RECOMMANDATIONS AUX **PLATEFORMES**

> **Affecter des ressources supplémentaires aux équipes chargées de garantir la sécurité et le respect des droits humains,** telles que les équipes responsables de la confiance et de la sécurité, les équipes chargées des droits humains ainsi que celles chargées de la politique de contenu. **Veiller à ce que ces équipes soient pleinement opérationnelles tout au long de l'année, et pas uniquement en période électorale ou dans d'autres contextes très sensibles.**

Réévaluer en permanence les politiques et les solutions techniques adoptées pour lutter contre le risque de tromperie numérique dans le domaine de la communication politique et d'autres situations à fort enjeu. Cet examen continu est essentiel pour garantir la conformité juridique et l'efficacité des stratégies de plateforme face à l'évolution rapide des technologies d'IA générative. Le tableau 2.2 propose quelques orientations générales, sans prétendre à l'exhaustivité.

205 Pour une discussion détaillée sur les recommandations politiques à mettre en place par les plateformes et imposées par les États lors des élections, voir «Protéger les élections démocratiques par la sauvegarde de l'intégrité de l'information» (2024), International IDEA, Forum sur l'information et la démocratie, Democracy Reporting International. Disponible sur : https://informationdemocracy.org/2024/01/30/tech-firms-governments-urged-to-combat-digital-election-threats/ (Consulté le : 8 février 2024).

206 Le rapport Le pluralisme de l'information dans les algorithmes d'indexation et de curation, publié par le Forum sur l'information et la démocratie en 2023, recommande, entre autres, d'« interdire aux campagnes politiques et aux acteurs politiquement affiliés de cibler (ou de ne pas cibler intentionnellement) des publics similaires avec des messages personnalisés » (p.23).

Tableau 2.2 Incidences de l'IA générative sur les formes habituelles de tromperie numérique utilisées dans la communication politique et dans d'autres contextes à fort enjeu, et les limites juridiques et les stratégies de remédiation possibles.

Formes habituelles de tromperie dans le domaine de la communication politique et d'autres contextes à fort enjeu	Utilisation de l'IA générative dans le cadre de ces pratiques	Limites juridiques[207]	Stratégies pour atténuer l'utilisation intentionnelle et involontaire des techniques d'IA générative
Mésinformation	Les systèmes d'IA générative peuvent involontairement entretenir la mésinformation politique, telle que les théories du complot, en raison de biais dans leurs données d'apprentissage, d'une mauvaise compréhension du contexte ou du potentiel d'hallucination des systèmes d'IA.	La légalité de la mésinformation est une question complexe qui s'entrecroise avec les considérations relatives à la liberté d'expression. Elle varie également en fonction de la juridiction et de la nature spécifique des éléments de mésinformation.[208]	• Mécanismes de signalement et de recours des utilisateurs. • Coopération avec des signaleurs de confiance. • Coupe-circuits viraux. • Vérification des faits, technologie humaine et IA. • Transparence des sources (telles que l'authenticité du contenu et les médias certifiés) - coopération avec des institutions de confiance. • Sensibilisation du public à l'identification et au signalement des fausses informations et de la désinformation, et à l'analyse critique du contexte des informations qu'il consulte. • Red-teaming et évaluation des risques. • Prise en compte des théories du complot dans l'apprentissage par renforcement par l'ajout d'étiquettes appropriées. • Réduction de l'utilisation de l'optimisation de l'engagement pour les contenus sensibles, y compris la politique et la santé.
Désinformation	L'IA générative peut contribuer à la création et à la diffusion de la désinformation avec une efficacité et une crédibilité accrues. Sa capacité à générer des textes, des images ou des vidéos sur mesure pour des publics spécifiques peut rendre les campagnes de désinformation plus ciblées et plus difficiles à identifier.	Certains pays proposent ou ont adopté une législation spécifique à la désinformation, qui peut cependant remettre en cause la liberté d'expression. Pour d'autres, les amendements proposés ou la base juridique de la lutte contre la désinformation reposent sur d'autres textes législatifs, tels que le Code pénal, le droit civil, le droit électoral, les lois régissant la diffamation ou l'incitation à la haine, le droit des télécommunications ou le droit de la cybersécurité.[209]	
Contenus média manipulés/ deepfakes et usurpation d'identité	Les technologies de l'IA générative réduisent considérablement le coût et facilitent le processus de création de discours, de déclarations ou d'actions synthétiques attribués à des personnalités ou à des autorités publiques.	La création et la diffusion de deepfakes politiques peuvent faire l'objet d'une action en justice en vertu du droit de la propriété intellectuelle et du droit de la protection de la vie privée. Dans certaines juridictions (p.ex. Californie, États-Unis[210] et État de Washington), les deepfakes font l'objet d'une législation spécifique. Certaines OSC américaines considèrent que les deepfakes utilisés dans les spots de campagne sont déjà couverts par la loi contre les « représentations frauduleuses »,[211] sans que la question ne soit tranchée[212] L'usurpation d'identité en ligne n'est pas directement considérée comme une infraction pénale,[213] mais elle peut relever de l'usurpation d'identité.[214] Elle peut également constituer un acte frauduleux.[215]	• L'utilisation de la technologie du deepfake devrait s'accompagner d'un consentement clair des personnes reproduites, et le contenu devrait être clairement étiqueté comme modifié ou généré par l'IA. Des réglementations spécifiques pourraient être nécessaires pour les œuvres satiriques et artistiques. • Utilisation de l'IA pour identifier et signaler les vidéos et les images relevant du deepfake. • Coopération avec des mécanismes de détection de pointe. • Intégration de marqueurs invisibles dans les médias et les contenus authentiques pour en vérifier l'originalité. • Application de la législation et de la réglementation contre la création et la distribution de deepfakes illégaux. • [La Californie et l'État de Washington interdisent les deepfakes dans une certaine période avant une élection, à moins que la communication n'indique de manière claire et concise que le deepfake a été généré artificiellement. Ils prévoient également des exemptions pour les satires ou parodies évidentes]. • Offrir la possibilité de vérification des utilisateurs de la plateforme. • Informer les utilisateurs des risques d'usurpation d'identité et de la manière de les repérer.

207 La légalité de ces pratiques varie considérablement d'un pays et d'une région à l'autre.
208 Funke, D. and Flamini, D. (2018). A guide to anti-misinformation actions around the world, Poynter. Disponible sur : www.poynter.org/ifcn/anti-misinformation-actions/ (Consulté le 8 février 2024).
209 Bontcheva, K. et al (2020). Legislative and Regulatory Responses to Disinformation, Excerpt from the Original Report, Broadband Commission for Sustainable Development. Disponible sur : https://en.unesco.org/sites/default/files/balanceact_legislative_en.pdf (Consulté le 8 février 2024).
210 Halm, K.C.,Kumar, A., Segal, J. and Kalinowski, C. (2020).Two New California Laws Tackle Deepfake Videos in Politics and Porn. Disponible sur : www.dwt.com/blogs/media-law-monitor/2020/02/two-new-california-laws-tackle-deepfake-videos-in (Consulté le 12 février 2024).
211 Public Citizen (2023). Comment to FEC: A.I.-Generated Political Deepfakes Are 'Fraudulent Misrepresentation'. Disponible sur : www.citizen.org/article/comment-to-fec-a-i-generated-political-deepfakes-are-fraudulent-misrepresentation/ (Consulté le 8 février 2024).
212 Fung, D.O., Brian (2023). First on CNN: Biden campaign prepares legal fight against election deepfakes, CNN Politics. Disponible sur : https://edition.cnn.com/2023/11/30/politics/biden-campaign-prepares-against-deepfakes/index.html (Consulté le 8 février 2024).
213 Bizga, A. (2020). *Blog Post: What is impersonation?,* Bitedefender. Disponible sur : www.bitdefender.com/blog/hotforsecurity/what-is-impersonation/ (Consulté le 8 février 2024).
214 Bitdefender (n.d.). *What is social media impersonation?*. Disponible sur : https://www.bitdefender.com/cyberpedia/what-is-social-media-impersonation/ (Consulté le 8 février 2024).
215 Bizga, A. (2020). *Blog Post: What is impersonation?,* Bitedefender. Disponible sur : www.bitdefender.com/blog/hotforsecurity/what-is-impersonation/ (Consulté le 8 février 2024).

Alarmisme et manipulation émotionnelle[216]	Les systèmes d'IA, qui intègrent l'analyse des signaux émotionnels et des modèles de langage, peuvent générer des contenus conçus pour susciter des réactions émotionnelles spécifiques, telles que la colère ou la peur, et manipuler ainsi l'opinion publique.	Si l'IA peut causer un préjudice ou inciter à la violence (ou est susceptible de la provoquer), elle peut être illégale (p.ex. le discours anxiogène à l'encontre des LGBTQ). Toutefois, les arguments émotionnels sont couramment utilisés dans la communication et les campagnes politiques.	• Application de mesures d'atténuation des risques sur les plateformes afin d'empêcher la diffusion de contenus générés par l'IA et susceptibles d'inspirer la peur. • Fournir des informations factuelles en coopération avec des fact-checkers et d'autres institutions indépendantes telles que les médias ou les autorités administratives afin de contrer les discours anxiogènes. • Recommandations spécifiques aux pandémies à l'intention des gouvernements217 : (1) des messages sereins et étayés scientifiquement de la part des autorités de santé publique ; (2) des avertissements d'interdiction adressés aux auteurs d'allégations extravagantes ou inappropriées ; (3) des actions en justice fermes et largement médiatisées contre les personnes et les entités formulant de fausses déclarations, afin de protéger un public rendu vulnérable par ses réactions émotionnelles face au phénomène de la pandémie. • Interdiction du ciblage des contenus basés sur des données personnelles sensibles.
Omissions	Les systèmes d'IA générative peuvent omettre certaines informations, intentionnellement ou non (p.ex. au moyen du fine tuning). Ce phénomène peut être particulièrement préoccupant si les systèmes d'IA sont utilisés pour la production d'informations.	L'omission délibérée de faits essentiels pourrait être contraire aux lois relatives à la communication honnête dans le cadre de la publicité et des campagnes politiques.	• Garantie de la pluralité du traitement de l'information. • Examen humain et responsabilité éditoriale. • Renforcement de la diversité et de la représentation au sein des systèmes d'IA via la provenance des données, l'étiquetage et le red-teaming.
Astroturfing	L'IA est en mesure d'automatiser la création de faux profils et de faux contenus, en simulant un soutien populaire à une cause ou à une opinion. Étant donné que l'IA générative peut apprendre à imiter les tendances ou les modèles de contenu, ce type de tromperie pourrait devenir plus difficile à détecter.		• Recours à l'IA pour identifier et supprimer les bots ou les faux comptes utilisés dans le cadre de campagnes d'astroturfing. • Application de réglementations exigeant la divulgation des contenus sponsorisés et des annonces politiques. • Mise en avant des contenus et des utilisateurs authentifiés. • Mise en place de mécanismes de signalement et d'examen par les utilisateurs. • Coopération avec des signaleurs de confiance et des fact-checkers.
Campagnes d'influence ciblant des communautés spécifiques	Les technologies d'IA destinées à la transposition du style et à l'adaptation du contenu peuvent aider à concevoir des campagnes d'influence trompeuses et manipulatrices, extrêmement ciblées et spécifiques à une communauté. Outre une plus grande efficacité, ces campagnes pourraient également être plus difficiles à identifier que les opérations d'influence étrangères traditionnelles, en particulier dans les régions non anglophones et les pays de « la majorité mondiale ». Ces capacités constituent une menace importante pour l'intégrité des élections, non seulement en termes de désinformation, mais également dans le contexte plus étendu du façonnement du discours politique, privant encore davantage de leurs droits des groupes déjà marginalisés et sous-représentés.		• Mise en place d'une infrastructure de contrôle de l'espace public informationnel au sein d'un pays, en particulier en période électorale. Cela implique de développer des outils permettant de reconnaître et de contrer les contenus fallacieux destinés à des communautés spécifiques, en privilégiant le respect de l'intégrité du débat politique dans des contextes linguistiques et culturels diversifiés. • Combler les lacunes en matière d'information par des informations fiables. • Interdiction des contenus ciblés basés sur des données personnelles sensibles.

216 Les pratiques d'IA interdites par la proposition de règlement de l'UE sur l'IA comprennent la mise en service d'un système d'IA qui aurait recours à des techniques subliminales, manipulatrices ou trompeuses, dans le but de modifier leur comportement (article 5.1).

217 Freckleton, I. (2020). COVID-19: Fear, quackery, false representations and the law, International Journal of Law and Psychiatry, Volume 72. Disponible sur : https://doi.org/10.1016/j.ijlp.2020.101611 (Consulté le 7 février 2024).

RECOMMANDATIONS
AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Effectuer des évaluations des risques et de red-teaming rigoureuses et exhaustives et instaurer des dispositifs d'examen accélérés pour les systèmes d'IA (tels que les chatbots) qui seront utilisés dans des contextes politiques et autres situations hautement sensibles** (voir chapitre 1, section 2).

> **Adresser aux utilisateurs des avertissements clairs et visibles sur le risque de mésinformation pouvant être générée par inadvertance par les chatbots. Ces avertissements devraient être accompagnés d'une recommandation de contre-vérification des informations à l'aide de sources fiables.**

4.3. TRAITEMENT DES RÉCLAMATIONS ET PROCÉDURES DE RECOURS

Les systèmes d'IA sont susceptibles de nuire à des individus et à des groupes, dans la mesure où ils peuvent produire des décisions et des résultats illégaux et préjudiciables. Afin de protéger et de faire respecter de manière proactive les droits des utilisateurs lors des interactions avec les systèmes d'IA et de créer des moyens de compensation et de recours, les déployeurs d'IA doivent intégrer des procédures de traitement des réclamations et des recours directement au sein de ces systèmes d'IA.

Les personnes originaires de pays dotés d'institutions fragiles pourraient être confrontées à des difficultés particulières pour intenter un recours devant les tribunaux nationaux contre des systèmes d'IA déployés à l'échelle internationale. Dans un tel contexte, ces procédures sont d'autant plus indispensables. Par conséquent, les organismes déployant des systèmes d'IA sont invités à mettre en place des mécanismes efficaces pour le dépôt de plaintes et l'obtention de réparations dans les cas où l'utilisation de l'IA enfreint les droits et la législation concernant, par exemple, la protection des données, la diffamation ou la violation des droits d'auteur, ou lorsqu'elle porte atteinte à la réputation.

Le risque d'atteinte à la réputation des organismes médiatiques par des systèmes d'IA générative qui attribueraient à tort des sources hallucinées ou qui fabriqueraient des articles semblant provenir de ces organismes est particulièrement préoccupant, d'où la nécessité de prévoir un traitement préférentiel des réclamations et une procédure de recours. Cette situation est particulièrement problématique à une époque où l'intégrité de l'information est fondamentale et où la différence entre les informations factuelles et les informations fabriquées peut influencer de manière significative l'opinion et les débats publics.[218]

Dans ce contexte, les tribunaux nationaux ont un rôle essentiel pour arbitrer les litiges liés au contenu généré par l'IA, en aidant à interpréter les lois nationales au regard de la technologie émergente et en prodiguant des conseils sur le caractère exécutoire des décisions à l'encontre des entités transnationales. La nomination d'un ombudsman de l'IA peut également faciliter la recherche d'une solution à l'amiable en cas de réclamation. L'interaction entre les mécanismes de recours internes des systèmes d'IA, les rôles d'arbitrage des tribunaux nationaux et un médiateur est examinée plus en détail au chapitre 4, sections 1.2 et 2.4.

218 South African Competition Commission (2023). Final Terms of Reference (ToR) for the Media and Digital Platforms Market Inquiry, Government Gazette No. 49309. Disponible sur : www.gov.za/sites/default/files/gcis_document/202309/49309gon3880.pdf (Consulté le 8 février 2024).

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Inclure des procédures de traitement des réclamations et de recours directement dans les systèmes d'IA afin de permettre aux utilisateurs de déposer plainte et de demander réparation dans les cas où l'utilisation d'un système d'IA enfreint les droits et la législation. Pour garantir une procédure exhaustive et équitable de traitement des plaintes, ces mécanismes de recours devraient être conformes aux sept principes énoncés dans les *Principes directeurs relatifs aux entreprises et aux droits de l'homme* des Nations unies pour des dispositifs efficaces de traitement des plaintes. Ces principes sont les suivants :**

- **Légitimité, c.-à-d. la nécessité que ces mécanismes bénéficient de la confiance des individus et des groupes qu'ils sont censés représenter et qu'ils soient responsables du traitement équitable des griefs.** Garantir la légitimité des systèmes d'IA équivaut à garantir que tout mécanisme de traitement des plaintes et de recours est conçu dans un souci de fiabilité et d'impartialité. Cela implique la transparence dans la prise de décision et l'absence de conflits d'intérêts susceptibles d'influencer l'issue d'une plainte.
- **Accessibilité, c.-à-d. la nécessité que ces mécanismes soient non seulement disponibles, mais également facilement compréhensibles et applicables par tous les utilisateurs, indépendamment de leur expérience et de leurs compétences.** Cela implique de s'assurer que les utilisateurs sont conscients de l'existence de ces mécanismes, qu'ils peuvent y accéder sans contrainte excessive et qu'ils bénéficient de l'assistance nécessaire pour suivre le processus. Il peut s'agir d'interfaces simples à utiliser, d'une assistance multilingue, d'instructions claires et de garanties de sécurité et de confidentialité pour les personnes qui portent plainte ou demandent réparation.
- **Prévisibilité, c.-à-d. la nécessité de fournir une procédure claire et reconnue, assortie d'un délai indicatif pour chaque étape.** Dans le contexte des systèmes d'IA, l'application du principe de prévisibilité suppose de veiller à ce que les utilisateurs comprennent le fonctionnement des mécanismes de réclamation et de recours. Les étapes du dépôt d'une plainte, le délai de traitement prévu et les résultats ou recours possibles doivent ainsi être clairement exposés.
- **Équité, c.-à-d. la nécessité de garantir que les utilisateurs plaignants disposent d'un accès raisonnable aux sources d'information, aux conseils et à l'expertise nécessaires pour s'engager dans une procédure de réclamation dans des conditions équitables, informées et respectueuses.** L'application de ce principe aux systèmes d'IA signifie que les utilisateurs doivent bénéficier d'un accès équitable aux ressources nécessaires pour répondre à leurs préoccupations. Cela pourrait impliquer de leur fournir des informations claires sur le fonctionnement du système d'IA, des conseils sur la procédure de plainte et la possibilité d'obtenir l'avis d'un expert. Il convient également de veiller à ce qu'il n'y ait pas d'obstacles à une procédure de réclamation efficace, tels qu'un jargon technique complexe ou des procédures de dépôt de plainte peu accessibles.

- **Transparence, c.-à-d. la nécessité pour les plaignants d'être tenus informés de l'évolution de leur dossier et de disposer d'informations suffisantes sur les performances du mécanisme pour avoir confiance en son efficacité.** S'agissant des systèmes d'IA, la transparence requiert une communication claire avec les utilisateurs sur la manière dont leurs plaintes spécifiques sont traitées et sur les critères utilisés pour les évaluer, dans le respect de la confidentialité. En outre, la transparence exige la publication de statistiques, d'études de cas ou d'informations plus détaillées sur l'efficacité du mécanisme dans le traitement des griefs.
- **Conformité aux droits, c.-à-d. la nécessité de garantir que les résultats et les recours offerts par les mécanismes de règlement des griefs sont conformes au droit, aux normes et aux principes internationaux relatifs aux droits humains.** Dans le contexte des systèmes d'IA, l'application de ce principe implique de veiller à ce que les recours offerts pour tout grief lié au système d'IA n'enfreignent pas les principes des droits humains. En outre, même lorsque les griefs ne sont pas initialement définis comme relevant des droits humains, le mécanisme de règlement des griefs du système d'IA doit les évaluer et les traiter en tenant compte de ces droits.
- **Engagement en faveur de l'apprentissage continu, c.-à-d. la nécessité d'analyser régulièrement les plaintes et d'en tirer des enseignements.** Dans le cadre des systèmes d'IA, les mécanismes de réclamation intégrés doivent non seulement traiter les problèmes du moment, mais encore évoluer en fonction des enseignements tirés de chaque cas. Cela implique de surveiller et d'examiner en permanence les types de plaintes reçues, d'en comprendre les problèmes sous-jacents et d'apporter les ajustements nécessaires au système d'IA afin d'éviter que des problèmes similaires ne se reproduisent à l'avenir. Il peut également s'agir de mettre à jour les politiques, les procédures ou les pratiques en fonction des enseignements tirés des plaintes. En outre, ces mécanismes devraient être accessibles aux particuliers et faciliter l'action collective le cas échéant.

> **Prévoir un canal spécifique pour les plaintes déposées par les organismes de médias et les défenseurs des droits humains. Celui-ci devrait donner la priorité aux questions liées à l'atteinte à la réputation causée par la mauvaise attribution de récits fabriqués ou de sources qui semblent être liées à ces organismes de médias.**

> **Garantir un examen rapide des plaintes, une action adéquate et un retour d'information au plaignant, et permettre le renvoi de l'affaire devant un tribunal national ou international ou un ombudsman, le cas échéant.**

> **Prendre rapidement des mesures pour répondre aux plaintes et ajuster le système** (voir chapitre 1, section 2.3), **ce qui peut inclure la suppression de données à caractère personnel du système et la mise en place de filtres de contenu** (voir chapitre 1, section 1.1e).

RECOMMANDATIONS AUX **ÉTATS**

> **Promulguer des lois sur la protection des consommateurs ou d'autres textes législatifs imposant aux systèmes d'IA la mise en place de procédures de traitement des plaintes. Ces procédures devraient comprendre les exigences suivantes :**

- **Traitement des plaintes dans les délais impartis, en donnant la priorité aux plaintes relatives aux potentielles infractions aux droits humains.**
- **Un canal spécial pour le traitement des plaintes des représentants des médias et des organismes de défense des droits humains.**
- **Publication de lignes directrices claires sur les procédures de traitement des plaintes, telles que les délais, les critères d'éligibilité et d'évaluation, les résultats escomptés et les recours possibles.**
- **Publication des décisions accompagnées des motifs et de références claires au droit international des droits humains et à la législation nationale correspondante** (pour plus d'informations sur les mécanismes de recours, voir le chapitre 4, section 2.4).

CHAPITRE 3 :
ENCOURAGER L'IA ÉTHIQUE

INTRODUCTION

Les réglementations imposant des mesures spécifiques pour favoriser le développement d'une IA éthique, bien qu'essentielles, ne représentent qu'un élément de l'équation. Pour mettre en place une culture de l'IA éthique, les responsables politiques et les autres parties prenantes concernés doivent réfléchir à des approches alternatives, conçues non seulement pour décourager les pratiques malveillantes, mais aussi encourager le développement et l'utilisation responsables de l'IA.

Responsabiliser les individus grâce à une éducation à l'IA est l'un des piliers de ce changement. En permettant à toutes les catégories de la population d'accéder aux connaissances et aux compétences nécessaires à la compréhension et à l'évaluation critique des systèmes d'IA, nous contribuons à mieux informer et impliquer le public, afin qu'il revendique des pratiques éthiques, valorise le développement responsable et fasse un usage responsable de ces outils.

Un autre fondement de ce changement consiste à créer un écosystème positif qui encourage le développement et le déploiement d'une IA éthique au moyen de codes de conduite volontaires, de certifications, de récompenses et d'incitations financières claires, telles que les marchés publics et les fonds publics. En outre, cet écosystème doit garantir que les ingénieurs et les personnes en charge du développement de l'IA possèdent une excellente connaissance des lois, normes et principes internationaux en matière de droits humains.

Enfin, il est essentiel de démocratiser le développement et l'accès à la technologie et aux systèmes d'IA afin d'éviter qu'une petite minorité d'entreprises ne contrôle l'espace informationnel. Cela suppose de créer des alternatives publiques aux systèmes d'IA à but lucratif, en veillant à ce que ces systèmes puissent être adaptés aux besoins et au contexte culturel des pays et des communautés du monde entier.

Encourager l'éducation à l'IA, proposer des incitations financières et éthiques et appliquer des réglementations ciblées sont autant de moyens grâce auxquels les États peuvent créer un futur où le développement et l'utilisation éthiques de l'IA sont la norme, et contribuent de manière positive à la société dans son ensemble.

1. ENCOURAGER UN DÉVELOPPEMENT ET UN DÉPLOIEMENT ÉTHIQUES

Les considérations éthiques relatives à l'IA sont essentielles à la construction d'un espace informationnel qui respecte les idéaux de liberté, de pluralisme et de démocratie.

Pour réaliser cette ambition, nous devons encourager le développement et le déploiement de systèmes d'IA qui respectent ces valeurs par l'intermédiaire de codes de conduite d'autorégulation, de certifications et de systèmes d'évaluation efficaces, de récompenses, ainsi que de fonds publics conditionnés au respect de normes éthiques strictes. En outre, nous devons mettre en valeur le potentiel des logiciels libres et des alternatives publiques aux systèmes d'IA à but lucratif, intégrer une formation à l'éthique destinée aux spécialistes des technologies de l'information et de l'IA, mettre en place un contrôle démocratique des entreprises d'IA et prévoir de réelles protections pour les lanceurs d'alerte.

Ces mesures exhaustives visent collectivement à garantir que le développement futur des systèmes d'IA ne se limite pas à concrétiser les priorités démocratiquement établies en matière d'IA, mais qu'il garantisse également le respect des droits fondamentaux.

1.1 CODES DE CONDUITE

En l'absence d'une législation exhaustive, les codes de conduite reposant sur l'autorégulation sont devenus un outil important pour les développeurs et les déployeurs d'IA, qui peuvent ainsi manifester volontairement leur engagement à respecter des normes éthiques. Ces codes, tels que le Code de conduite du G7,[219] de préférence élaborés dans le cadre d'un processus participatif, fournissent un cadre d'orientation précieux pour le développement et le déploiement éthiques de l'IA. De tels codes ont également été intégrés dans la législation pour orienter le comportement des entreprises, et sont considérés comme faisant partie des mesures d'atténuation des risques.[220]

Néanmoins, le défi majeur consiste à transformer ces textes ambitieux en outils vraiment efficaces. Il est essentiel de mettre en place des mécanismes d'application rigoureux et de surmonter les obstacles à la mise en œuvre pour exploiter pleinement le potentiel des codes d'autorégulation et faire progresser les pratiques responsables en matière d'IA.

219 Commission Européenne (2023).Code de conduite international pour les systèmes d'IA avancés dans le cadre du processus Hiroshima. Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Consulté le 8 février 2024).

220 Le règlement de l'UE sur les services numériques charge la Commission européenne d'encourager les plateformes en ligne à élaborer des codes de conduite volontaires (article 45), en insistant sur le code de conduite pour la publicité en ligne (article 46) et l'accessibilité (article 47).

RECOMMANDATIONS AUX **ÉTATS**

> **Prendre en compte et favoriser le rôle des codes de conduite reposant sur l'autorégulation à titre d'approche complémentaire de la réglementation, y compris au moyen d'une assistance technique, en fournissant par exemple des conseils sur l'élaboration de ces codes d'une manière inclusive.**[221]

> **Développer un système de récompense pour les entreprises qui se conforment à des codes d'autorégulation reconnus.** Dans ce cadre, **il est possible d'envisager d'associer les opportunités de marchés publics à l'adhésion à des codes d'autorégulation reconnus** (voir section 1.4). Cette démarche peut encourager davantage l'engagement et favoriser les meilleures pratiques dans l'ensemble du secteur.

> **Mettre en place des dispositifs clairs permettant aux développeurs et aux utilisateurs de l'IA de signaler les infractions aux principes énoncés dans ces codes et offrir une protection aux lanceurs d'alerte** (voir section 1.8).

> **Associer expressément l'adhésion aux codes de conduite à la responsabilité. À cet égard, l'adhésion peut être considérée comme l'une des mesures de gestion des risques requises** (voir chapitre 2, section 1.1).

> **Imposer aux entreprises de communiquer publiquement leur adhésion aux codes de conduite, ainsi que leurs pratiques de conformité, et encourager la mise en place de mécanismes d'examen de ces rapports de conformité en soutenant financièrement des organismes de recherche indépendants et des organisations de la société civile** (voir chapitre 4, sections 1.4 et 3.2). Afin de faciliter le contrôle et de favoriser la transparence dans l'évaluation du respect des codes de conduite par les entreprises, les États sont invités à promouvoir la mise en place d'une base de données centralisée, en libre accès et facile à utiliser (voir chapitre 4, section 4.1).

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Adhérer à des codes de conduite reposant sur l'autorégulation reconnus par des organismes du secteur réputés ou par les autorités réglementaires compétentes.**

> **Collaborer avec différents partenaires, notamment des chercheurs, des organisations de la société civile et des organismes gouvernementaux compétents, pour veiller à ce que les principes formulés dans les codes de conduite soient exhaustifs, pertinents et conformes aux attentes et aux valeurs de la société, ainsi qu'à l'intérêt public général.**

> **Transformer concrètement les principes des codes de conduite en étapes pratiques et les intégrer dans les processus de développement et de déploiement.**

221 La proposition de règlement de l'UE sur l'IA suggère l'élaboration de codes de conduite destinés à favoriser l'application volontaire aux systèmes d'IA autres que les systèmes d'IA à haut risque (article 69).

> **Communiquer publiquement l'adhésion aux codes d'autorégulation et démontrer l'engagement en faveur d'une IA éthique en publiant régulièrement des rapports sur les pratiques et les progrès de l'entreprise. Ces rapports devraient servir à indiquer précisément dans quelle mesure et par quelles actions les entreprises respectent leurs engagements, et quels sont les domaines où des améliorations sont possibles.**

1.2 CERTIFICATIONS ET ÉVALUATIONS

La pression sociale et l'attention que le public accorde aux pratiques éthiques dans d'autres secteurs peuvent également concerner l'IA. Dans ce contexte, un système de certification adapté aux entreprises et aux entités d'IA, comparable à la certification du commerce équitable[222] pour les pratiques commerciales éthiques, pourrait contribuer à encourager un développement et un déploiement responsables. Ce système permettrait de distinguer clairement et de manière vérifiable les entreprises qui s'engagent à respecter des principes éthiques en matière d'IA, et aiderait les consommateurs à prendre des décisions éclairées.

Bien qu'un tel processus de certification devrait être très exhaustif, une première étape d'évaluation du respect des normes par les systèmes d'IA pourrait mesurer la transparence, la sécurité ou l'éthique telles que définies par la société civile, des chercheurs ou les autorités chargées de l'IA. L'indice de transparence du modèle de la fondation de l'université de Stanford[223] en est un parfait exemple : les indicateurs pertinents qu'il fournit aux utilisateurs, pouvoirs publics et autres parties prenantes peuvent être pris en considération dans les règles de passation des marchés, ce qui crée une concurrence saine entre les développeurs et les fournisseurs de systèmes d'IA, qui les incite à améliorer leurs systèmes.

RECOMMANDATIONS AUX **CSOS**

Encourager la création d'un système de certification sur mesure pour les entreprises d'IA, inspiré du succès du système de certification du commerce équitable. Cela suppose :

> **L'élaboration d'un ensemble complet de normes éthiques pour le développement et le déploiement de l'IA afin de consolider la certification.**

> **La création d'un organisme indépendant à but non lucratif, ou le recours à un organisme existant, pour gérer le processus de certification.** Cet organisme serait chargé d'accréditer les agences de certification, de définir et de faire respecter les critères d'évaluation et de délivrer des certifications aux entreprises concernées.

> **La promotion de la participation active des différentes parties prenantes de l'écosystème de l'IA, y compris les développeurs, les déployeurs, les utilisateurs, les chercheurs, les journalistes et les organisations de la société civile, en insistant sur la diversité géographique, linguistique, culturelle et cognitive.** Cela garantirait que des perspectives plurielles, y compris celles de « la majorité mondiale », influencent la conception et la mise en œuvre du système de certification.

222 Fair Trade International (n.d.). *About Us*. Disponible sur : www.fairtrade.net/about (Consulté le 9 février 2024).
223 Miller, K. (2023). Introducing The Foundation Model Transparency Index, Stanford University. Disponible sur : https://hai.stanford.edu/news/introducing-foundation-model-transparency-index (Consulté le 9 février 2024).

> L'organisation de campagnes de sensibilisation du public afin d'instruire les utilisateurs à l'importance d'une IA éthique et à la signification du système de certification.

> **La mise en place d'un mécanisme d'audit pour s'assurer que les entreprises se conforment systématiquement aux recommandations.** Les entreprises jugées non conformes se verraient retirer cette certification.

> **La création d'un registre public qui permettrait un audit public et renforcerait la transparence et la confiance.**

> **La garantie de financement en exigeant des entreprises et entités du secteur de l'IA qu'elles s'acquittent d'une redevance pour le processus de certification.**

RECOMMANDATIONS AUX **ÉTATS**

> **Encourager la participation à un système de certification de type « commerce équitable » et l'obtention d'un score élevé en matière de transparence, de sécurité ou d'éthique, en les prenant en compte dans les décisions financières relatives aux entreprises et entités de l'IA.**

> **Envisager la mise à disposition de ressources financières pour financer la création et la gestion d'un tel système de certification de type « commerce équitable ».**

1.3. RÉCOMPENSES

La reconnaissance publique qui accompagne les programmes de récompenses aux niveaux national, régional ou international peut inciter le développement et le déploiement d'une IA éthique. Un tel programme doit s'appuyer sur des critères de sélection clairs, équitables et publics, et faire appel à un jury indépendant. Les Digital Government Awards des Nations unies sont un exemple d'une telle approche.[224]

RECOMMANDATIONS AUX **ÉTATS**

> **Mettre en place un programme de récompenses visant à distinguer et à récompenser les réalisations exceptionnelles en matière de développement et de déploiement d'une IA conforme à l'éthique[225], incluant de possibles avantages fiscaux et/ou des récompenses non financières telles que la reconnaissance publique.**

224 CNUCED (2023). UN Digital Government Awards celebrate excellence in online public services. Disponible sur : https://unctad.org/news/un-digital-government-awards-celebrate-excellence-online-public-services (Consulté le 9 février 2024).

225 Ces prix devront répondre à des critères d'attribution clairs. Ils pourraient être attribués à différents projets d'IA tels que ceux gérés par des PME locales, des organisations à but non lucratif, des instituts de recherche ou la société civile.

1.4. MARCHÉS PUBLICS ET FONDS PUBLICS

Les marchés publics, qui représentent environ 12,9 % du PIB dans les pays de l'OCDE,[226] 17 % en moyenne dans les pays africains[227], 6 % en Amérique latine et dans les Caraïbes[228], et jusqu'à 20 % dans les pays de l'ANASE[229], offrent une occasion sans pareille de définir le marché de l'IA. En appliquant des normes éthiques strictes en matière d'IA dans les processus de passation de marchés et en privilégiant les développeurs et déployeurs d'IA responsables, les gouvernements peuvent orienter l'industrie vers des pratiques éthiques. Une telle approche permet non seulement aux gouvernements d'acquérir des technologies de pointe, mais également de montrer l'exemple. De plus, l'attribution de fonds et d'investissements publics devrait être conditionnée au respect de normes éthiques par les entreprises et entités d'IA, ce qui garantirait que les finances publiques contribuent au développement de systèmes d'IA conformes à l'intérêt général. Les avantages fiscaux peuvent constituer une mesure d'encouragement supplémentaire au développement et au déploiement de systèmes d'IA éthiques.

En soutenant l'implication directe des États dans la promotion d'un développement éthique de l'IA, il est essentiel de reconnaître que, dans certains contextes, les institutions sont susceptibles d'être instrumentalisées par des intérêts privés pouvant exercer une influence notable sur les politiques locales, y compris en matière de passation de marchés. Pour se prémunir contre toute influence privée inappropriée, il est indispensable de mettre en place des procédures de passation de marchés publics solides et transparentes. Celles-ci doivent être transparentes et ouvertes à tous, et comporter des garanties telles que des exigences de transparence (par exemple la divulgation publique des appels d'offres, des portails publics permettant aux parties extérieures de vérifier l'attribution des marchés et les critères utilisés pour évaluer cette attribution) et des politiques relatives aux conflits d'intérêts.[230]

RECOMMANDATIONS AUX **ÉTATS**

> **Élaborer des lignes directrices claires sur les types de systèmes d'IA qui devraient être conçus, achetés et utilisés dans le secteur public, et dans lesquels il conviendrait d'investir. Ces lignes directrices devraient définir les règles et réglementations auxquelles les systèmes d'IA doivent se conformer :**
>
> - Dans la mesure du possible, privilégier les alternatives publiques aux systèmes à but lucratif, les systèmes open-source ou les systèmes d'IA éthiques plutôt que les solutions propriétaires non transparentes.[231]
> - Dans la mesure du possible, privilégier les systèmes d'IA certifiés par la certification « commerce équitable » et bénéficiant d'un score élevé en matière de transparence, de sécurité ou d'éthique.

226 OCDE (2024), Poids des marchés publics, Panorama des administrations publiques 2023. Disponible sur : https://www.oecd-ilibrary.org/fr/governance/panorama-des-administrations-publiques-2023_e5bad4fb-fr (Consulté le 9 février 2024).
227 Arisoy, E., Leipold, K. and Messan, K. (2023). *The expanding role of public procurement in Africa's economic development*, World Bank Blogs. Disponible sur : https://blogs.worldbank.org/governance/expanding-role-public-procurement-africas-economic-development (Consulté le 9 février 2024).
228 OCDE (2020), Government at a Glance: Latin America and the Caribbean 2020. Disponible sur : www.oecd.org/publications/government-at-a-glance-latin-america-and-the-caribbean-5ceda53e-en.htm (Consulté le 9 février 2024).
229 PNUD. Improving Procurement Transparency. Disponible sur : www.undp.org/asia-pacific/fairbiz/improving-procurement-transparency (Consulté le 9 février 2024).
230 OCDE (2015). *Recommandation du Conseil sur les marchés publics*. Disponible sur : https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0411(Consulté le 9 février 2024).
231 La Loi française sur le Numérique (Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique) encourage l'utilisation de logiciels open-source pour les administrations publiques développant, achetant ou utilisant des logiciels (Article 16). Disponible sur : www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000033203039 (Consulté le : 9 février 2024).

> **Faire un usage stratégique des finances publiques, des déductions fiscales, des crédits et des exonérations pour favoriser le développement et le déploiement de systèmes d'IA éthiques et locaux.**

1.5 SOLUTIONS ALTERNATIVES OPEN-SOURCE ET PUBLIQUES AUX SYSTÈMES D'IA À BUT LUCRATIF

Dans la mesure où les systèmes d'IA jouent un rôle de plus en plus important dans le progrès technologique et la croissance économique, l'accès à cette technologie devient primordial pour encourager l'innovation et garantir une participation équitable. Les alternatives publiques aux systèmes d'IA à but lucratif et les systèmes open-source sont des stratégies incontournables pour relever ces défis et bâtir un avenir plus inclusif en matière d'IA.

L'IA open-source (y compris une licence open-source pour le code, les données, ainsi que les poids du système[232] et la publication complète de ces trois composantes) offre une occasion unique de démocratiser l'accès à des technologies très performantes et d'encourager l'innovation tout en prônant la transparence et la collaboration. L'objectif est notamment de permettre la conception et le déploiement de systèmes d'IA détenus et gérés par des entités publiques, ainsi que par la société civile, le secteur de la recherche, les organismes de médias et les acteurs privés. Cependant, les préoccupations concernant les vulnérabilités en matière de sécurité et l'utilisation abusive potentielle par des parties malveillantes exigent une approche prudente dans ce paysage complexe. Des études ont ainsi indiqué que les systèmes intégrés d'atténuation des risques ou de filigrane peuvent être facilement supprimés par des acteurs malveillants.[233]

Par ailleurs, les États devraient fournir des ressources financières pour soutenir le développement d'infrastructures numériques publiques et d'alternatives publiques aux systèmes d'IA à but lucratif, susceptibles de rééquilibrer la concentration du marché et d'offrir un accès à des systèmes sûrs et fiables, garantissant ainsi la protection des droits humains et le respect de normes éthiques. L'importance d'une telle infrastructure et de l'accès pour les organismes de médias est également soulignée dans les lignes directrices du Conseil de l'Europe.[234] Cette infrastructure publique comprend l'accès aux ensembles de données d'apprentissage des systèmes d'IA et la puissance de calcul pour permettre un accès plus démocratique au développement des systèmes d'IA, mais également au développement des systèmes d'IA, tels que les systèmes de recommandation ou les modèles de base. Ces alternatives publiques aux systèmes d'IA à but lucratif peuvent être confiées à des organismes administratifs indépendants, à des institutions similaires aux médias de service public ou à d'autres parties prenantes agissant dans l'intérêt public, telles que la société civile ou la communauté scientifique. Bien que ces systèmes soient de préférence publiés sous une licence libre, ils peuvent également être fermés (closed source).

232 Sijbrandij, S. (2023). AI weights are not open «source», Open Core Ventures. Disponible sur : https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/ (Consulté le 9 février 2024).
233 Zhang, H. et al. (2023). *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models.* Disponible sur : https://arxiv.org/abs/2311.04378 (Consulté le 15 février 2024).
234 Conseil de l'Europe (2023). Lignes directrices sur la mise en œuvre responsable de systèmes d'intelligence artificielle dans le journalisme. Point 6.2. Disponible sur : https://rm.coe.int/cdmsi-2023-014-lignes-directrices-sur-la-mise-en-uvre-responsable-de-s/1680adb4c7 (Consulté le 7 février 2024).

RECOMMANDATIONS AUX **ÉTATS**

> **Prendre acte de la double nature de l'IA open-source,** en reconnaissant à la fois son potentiel bénéfique et son potentiel néfaste. Cette perspective équilibrée encourage un développement et un déploiement responsables tout en optimisant l'impact positif de cette puissante technologie. Il s'agit notamment de :
  - ♦ Veiller à ce que les mêmes obligations, droits et résultats escomptés s'appliquent à tous les développeurs et déployeurs de systèmes d'IA, qu'il s'agisse de logiciels libres, d'alternatives publiques à des systèmes à but lucratif ou de systèmes propriétaires, dans la mesure du possible et dans le respect de la législation.
  - ♦ Prendre part à des discussions inclusives avec les parties prenantes concernées, y compris la société civile et la communauté universitaire, afin d'élaborer des mécanismes favorisant la recherche d'intérêt public et le développement de l'IA sur la base de systèmes open-source, et en tirant parti des valeurs de l'open-source, notamment la collaboration et la transparence. Cet objectif devrait être poursuivi parallèlement à la mise en place de garde-fous visant à protéger ces systèmes contre les utilisations malveillantes.

> **Instaurer un programme dédié au développement et à la gestion des jeux de données publiques de formation[235].** Ces derniers devraient être représentatifs de la diversité des populations nationales, de la plus haute qualité et facilement accessibles aux chercheurs et aux développeurs du monde entier, en particulier dans les pays de « la majorité mondiale ».
  - ♦ **Créer des cadres de collaboration internationale pour le partage des données et de l'expertise sur les jeux de données publiques de formation.** Cela favorisera l'échange de connaissances, le renforcement des capacités dans les régions à faibles ressources et l'élaboration de normes mondiales pour la collecte et l'utilisation éthiques des données.

> **Envisager des financements publics pour soutenir le développement et la gestion d'une infrastructure publique pour des systèmes d'IA dignes de confiance. Cette infrastructure pourrait comprendre des alternatives publiques aux systèmes de recommandation[236], de modération de contenu ou de classification à but lucratif, ainsi que des alternatives publiques aux systèmes d'IA générative à but lucratif.** Elle pourrait notamment concerner les systèmes d'IA utilisés dans l'espace informationnel et par les organismes de presse afin de démocratiser l'accès à des systèmes fiables. Ces derniers devraient de préférence être disponibles sous une licence open-source ou une licence éthique.[237]

> **Investir dans le développement d'une infrastructure d'espace informationnel numérique alternative permettant d'optimiser les résultats sociétaux, le niveau de confiance et le caractère démocratique et fiable de l'espace de l'information et de la communication**. Il peut s'agir (via des financements) de créer des infrastructures publiques pour soutenir l'émergence d'espaces d'information et

235 Dans le même esprit, la proposition de règlement de l'UE sur l'IA dispose que « l'espace européen des données de santé facilitera l'accès non discriminatoire aux données de santé et l'entraînement d'algorithmes d'intelligence artificielle à l'aide de ces jeux de données » et encourage les autres autorités compétentes sectorielles à faire de même (point 45).

236 Forum sur l'information et la démocratie (2023). Pluralisme de l'Information dans les Algorithmes d'Indexation et de Curation. Disponible sur : https://informationdemocracy.org/fr/pluralisme/ (Consulté le 8 février 2024).

237 Une licence éthique inclut les quatre libertés du code source ouvert tout en imposant des restrictions sur certains comportements, comme dans les licences d'IA responsable (s.d.). Disponible sur www.licenses.ai/ (Consulté le : 8 février 2024).

de communication alternatifs, gérés par des organisations communautaires et de la société civile, ou inspirés des modèles de médias de service public. Il peut également s'agir de financer des recherches visant à tester le fonctionnement de ces espaces (effets des mesures d'engagement, des systèmes d'authentification et des identités numériques, encouragement de comportements alternatifs tels que le dialogue, etc. Voir à ce propos le chapitre 1, section 1.4). Ces derniers devraient de préférence être disponibles sous une licence open-source ou une licence éthique.[238]

> **Consacrer des ressources au développement, à la gestion et à la mise à niveau de l'infrastructure numérique publique essentielle, ce qui peut comprendre l'accès à l'internet à haut débit, les centres de données, les plateformes d'informatique dans le cloud et une puissance de calcul adéquate.** Ces mesures permettront de jeter les bases d'une plus grande accessibilité à l'IA et de créer une demande pour des systèmes d'IA qui répondent aux besoins des utilisateurs actuellement privés de leurs droits. **Ces initiatives ne doivent en aucun cas conduire à des mandats de localisation des données susceptibles de mettre en péril les droits humains et/ou d'entraver la rentabilité.**

1.6 FORMATION À L'ÉTHIQUE POUR LES SPÉCIALISTES DES TI ET DE L'IA

L'éducation, y compris les programmes universitaires, joue un rôle crucial dans la définition des normes suivant lesquelles les développeurs de systèmes d'IA construiront les systèmes. Pour favoriser un développement et un déploiement éthiques, les États doivent encourager l'intégration de modules sur les principes éthiques et les responsabilités des spécialistes des technologies de l'information quant à la conception d'une IA éthique aux systèmes éducatifs. Cette mesure devrait également inclure une formation sur les implications, y compris les implications potentiellement néfastes des systèmes d'IA pour l'espace inforrmationnel. Une telle formation devrait être accessible tout au long de la carrière des spécialistes des technologies de l'information et de l'IA.

En outre, il est indispensable de veiller à la diversité des matières enseignées aux spécialistes des TI et de l'IA, afin de garantir la représentativité et l'inclusivité des systèmes d'IA. Cela implique nécessairement d'inclure les lois sur les droits humains, les études culturelles, l'histoire, la sociologie, etc., afin de s'assurer que les ingénieurs et autres spécialistes participant au développement et au déploiement de l'IA soient bien formés et disposent d'une compréhension globale des diverses incidences sociétales que leur travail pourrait avoir.[239]

Les entreprises et entités spécialisées dans l'IA doivent également investir dans la formation et les ressources destinées à leurs propres développeurs, en les dotant des connaissances et des outils nécessaires à l'identification et à l'atténuation des risques liés à l'IA. Cela implique une prise de conscience des diverses répercussions sociétales, une compréhension des lois, normes et principes relatifs aux droits humains, l'identification et la rectification des biais, ainsi que la conception proactive de systèmes conformes aux principes éthiques et aux valeurs démocratiques.

238 Une licence éthique inclut les quatre libertés du code source ouvert tout en imposant des restrictions sur certains comportements, comme dans les licences d'IA responsable (s.d.). Disponible sur www.licenses.ai/ (Consulté le : 8 février 2024).
239 Webb, A. (2019). The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity, PublicAffairs New York.

RECOMMANDATIONS AUX **ÉTATS**

> **Promouvoir la diversité des matières dans les programmes d'études des spécialistes des TI et de l'IA, y compris le droit relatif aux droits humains, les études culturelles, l'histoire et la sociologie.**

> **Inclure des modules sur l'éthique, les normes éthiques, les implications des systèmes d'IA pour l'espace informationnel et l'élaboration des politiques publiques dans la formation technique des développeurs d'IA** (cours universitaires, écoles spécialisées, etc.) **et proposer une formation continue sur ces questions.** À cette fin, les États devraient :

- Collaborer avec les partenaires concernés, y compris les établissements d'enseignement supérieur, des spécialistes de l'éducation, la société civile, des chercheurs, des universitaires et d'autres experts compétents, afin de concevoir ces programmes de formation et ces cours universitaires.
- Veiller à ce que ces modules se concentrent en particulier sur la manière dont les systèmes d'IA utilisés dans l'espace de l'information et de la communication peuvent entraver les droits humains et les libertés fondamentales, y compris, mais sans s'y limiter :
  - La liberté d'expression
  - Le pluralisme des médias
  - La représentation de cultures et de langues diverses
  - L'égalité d'accès aux connaissances artistiques, scientifiques et technologiques
  - Le droit à la vie privée
  - Le droit à la non-discrimination, qui comprend, sans s'y limiter, la race, la couleur, l'identité et l'expression de genre, l'orientation sexuelle, la langue, la religion, les opinions politiques ou autres, l'origine nationale ou sociale, la propriété et la naissance
  - L'accès à l'information et à des sources fiables
  - La liberté de la presse

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Inscrire la formation à l'éthique de l'IA comme un fondement du processus d'intégration des développeurs d'IA.**

> **Proposer au personnel des entreprises des modules spécifiques de formation continue sur l'éthique, les normes éthiques et les implications des systèmes d'IA pour l'espace informationnel.**

1.7 CONTRÔLE DÉMOCRATIQUE DES ENTREPRISES D'IA

Dans la plupart des entreprises d'IA, les membres du conseil d'administration et les investisseurs décident souvent des orientations stratégiques de l'entreprise ou de l'entité. Les intérêts commerciaux (bénéfices et valeur actionnariale) ont jusqu'ici généralement prévalu sur l'intérêt public et sur le développement et le déploiement éthiques des systèmes d'IA. Afin de garantir que l'intérêt public oriente les décisions stratégiques des entreprises d'IA, celles-ci devraient être mandatées pour mettre en place des structures de gouvernance ou de contrôle démocratiques et participatives.

Malgré certaines lacunes, le comité de surveillance de Meta[240] est un premier pas dans la bonne direction, car il permet un examen indépendant des décisions en matière de modération de contenu. À long terme, ces structures devraient être mandatées pour exercer un contrôle dépassant les décisions spécifiques, ce qui leur permettrait de participer aux décisions stratégiques et à l'élaboration de la politique de l'entreprise.

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux entreprises et entités du secteur de l'IA de mettre en place des structures de gouvernance démocratiques.**
>
> - Dans le cas où l'entreprise ne relève pas de la juridiction du pays, les procédures d'octroi de licences et les dispositions relatives aux marchés publics peuvent être utilisées pour encourager ce type de structures.
> - Fournir des conseils sur la mise en place de ces mécanismes. Ces derniers peuvent adopter différentes formes, telles qu'un conseil de surveillance, une assemblée citoyenne ou une représentation des employeurs et des utilisateurs.
>   - Les utilisateurs devraient pouvoir voter pour des représentants des utilisateurs habilités à faire des suggestions, à être consultés et à opposer leur veto aux décisions qui les concernent directement, telles que la protection et l'utilisation des données. Ces représentants siégeraient au conseil d'administration ou au conseil de surveillance.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Instaurer une gouvernance démocratique de l'entreprise ou de l'entité. En fonction des structures juridiques et des pratiques de gouvernance d'entreprise, cet objectif peut être atteint en :**
>
> - **Confiant un siège au conseil d'administration à des membres du public indépendants de l'entreprise et représentant les intérêts de la société, ainsi qu'à des représentants des usagers.** Cela leur permettrait d'avoir une visibilité sur les opérations de l'entreprise et l'autorité nécessaire pour examiner ses activités, en veillant à ce qu'elle œuvre dans le meilleur intérêt du public.[241]

240 Oversight Board. Disponible sur : www.oversightboard.com/ (Consulté le 9 février 2024).
241 Milmo, D. (2023). AI firms 'should include members of public on boards to protect society', The Guardian. Disponible sur : www.theguardian.com/technology/2023/dec/06/ai-firms-should-include-members-of-public-on-boards-to-protect-society (Consulté le 9 février 2024).

- **Créant un conseil de surveillance composé d'experts extérieurs indépendants et diversifiés et de représentants des utilisateurs.** Ce conseil de surveillance devrait fonctionner comme un organe distinct qui examinerait les décisions de l'entreprise concernant les systèmes d'IA, et fournirait des conseils supplémentaires à ce sujet. Le conseil devrait également avoir le pouvoir d'annuler ou d'émettre des décisions contraignantes. Par ses travaux, il devrait contribuer au respect de la législation internationale en matière de droits humains et promouvoir l'utilisation éthique et responsable de l'IA. [242]

Dans les deux cas, les entreprises et entités spécialisées dans l'IA doivent veiller à ce que les critères de sélection des membres soient équitables et transparents, afin de garantir la représentation d'une grande diversité de groupes, en particulier les plus susceptibles d'être affectés par les systèmes d'IA et les plus vulnérables à leur égard.

1.8 PROTECTION DES LANCEURS D'ALERTE

Les lanceurs d'alerte peuvent fournir des informations précieuses sur le non-respect des cadres de gouvernance et de responsabilité obligatoires, ainsi que sur les dysfonctionnements des systèmes d'IA, en communiquant des informations confidentielles aux autorités compétentes. Ils peuvent participer à des investigations ou à des procédures judiciaires menées par des entités qualifiées qui enquêtent sur des activités, des politiques, des pratiques ou des tâches définies considérées comme une infraction potentielle ou présumée aux lois, règles ou réglementations en vigueur.[243] Néanmoins, les employés actuels et anciens peuvent être réticents à exposer ou à rendre publics de tels manquements ou dysfonctionnements par crainte de représailles (p.ex. licenciement direct ou indirect, rétrogradation, suspension, menaces, harcèlement, inscription sur une liste noire ou toute autre mesure discriminatoire ou préjudiciable au personnel)[244] et en raison de l'absence de mesures d'incitation et de protections adéquates.

RECOMMANDATIONS AUX **ÉTATS**

> **Mettre en place des protections juridiques efficaces pour les lanceurs d'alerte anciens ou actuels employés de l'industrie de l'IA.** Celles-ci comprennent :
>
> - **La mise en place d'un dispositif permettant aux lanceurs d'alerte d'intenter une action en justice s'ils font l'objet de représailles à la suite de la divulgation d'infractions potentielles aux lois, règles et réglementations en vigueur ou d'autres comportements potentiellement contraires à l'éthique. Cela concerne également les divulgations contraires à l'éthique mais non illégales.**[245] Ce droit d'action privé devrait prévoir des recours et des protections appropriés.

242 Kulick, A. (2022). Meta's Oversight Board and Beyond – Corporations as Interpreters and Adjudicators of International Human Rights Norms, The Law and Practice of International Courts and Tribunals 2022, Forthcoming. Disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4226521 (Consulté le 9 février 2024).

243 Schakowky, J. (2021). A Bill to provide incentives for and protect whistleblowers under the authority of the Federal Trade Commission, and for other purposes. Disponible sur : https://schakowsky.house.gov/sites/evo-subsites/schakowsky-evo.house.gov/files/SCHAKO_082_xml.pdf (Consulté le 9 février 2024).

244 US Department of Labor. Retaliation. Disponible sur : www.whistleblowers.gov/know_your_rights (Consulté le 9 février 2024).

245 Frances Haugen, lanceuse d'alerte de Meta (anciennement Facebook), a révélé des comportements au sein de l'entreprise qui, sans être forcément illégaux, ont été jugés contraires à l'éthique par de nombreux observateurs.

- **L'établissement d'une responsabilité pénale** pour le harcèlement et l'intimidation des lanceurs d'alerte.
- **La mise en place de mesures visant à assurer la sécurité physique des lanceurs d'alerte,** notamment la sécurité personnelle, via des mesures de sécurité et de soutien en cas de stress ou de traumatisme.
- **L'interdiction des accords de confidentialité qui empêchent ou limitent la divulgation d'informations factuelles relatives à des infractions potentielles aux lois, règles, réglementations, normes éthiques ou droits humains dans le domaine de l'IA.**

> **Mettre en place une plateforme spéciale permettant aux lanceurs d'alerte de déposer leurs plaintes, y compris de manière anonyme, afin de les faire examiner rapidement.** Cette plateforme pourrait être gérée par l'autorité chargée de l'IA (voir chapitre 4, section 1.2).

> **Mettre en place des dispositifs de conseil juridique spéciaux et confidentiels pour les lanceurs d'alerte, leur conférant le droit d'être représentés par des avocats spécialisés dans la protection des lanceurs d'alerte.**[246]

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Instaurer un mécanisme interne permettant aux employés de déposer des plaintes, y compris de manière anonyme, afin de les faire examiner rapidement. Ces plaintes doivent être immédiatement portées à l'attention de la direction générale. Si le problème n'est pas traité, des poursuites judiciaires doivent être engagées à son encontre, pouvant aller jusqu'à des plaintes pour négligence en vertu du droit de la responsabilité** (voir chapitre 2, section 3.1). Toutes les entités qui reçoivent des informations divulguées par un lanceur d'alerte doivent en garantir la confidentialité.

> **Informer les employés des droits des lanceurs d'alerte et des conséquences juridiques de toute forme de représailles à l'encontre de ces derniers.**

> **Mettre en place des programmes de lutte contre les représailles, notamment en désignant un responsable indépendant au sein de l'entreprise, qui examinerait les rapports et les préoccupations des employés.**[247]

246 Marzotto, M. (2023). Whistleblowers are human rights defenders. So why don't we protect them like they are? The Signals Network. Disponible sur : https://thesignalsnetwork.org/whistleblowers-are-human-rights-defenders/ (Consulté le 9 février 2024).
247 US Department of Labor (n.d.). How to Create an Anti-Retaliation Program. Disponible sur : www.whistleblowers.gov/antiretaliation (Consulté le 9 février 2024).

2. ENCOURAGER UNE UTILISATION RESPONSABLE

Les systèmes d'IA ont littéralement révolutionné la manière dont l'information est générée, partagée, consommée et contrôlée. Du reportage automatisé à l'analyse sophistiquée des données et à la personnalisation du contenu, les outils d'IA ont la capacité de renforcer les pratiques journalistiques, d'améliorer l'interaction entre les gouvernements et les citoyens, et d'autonomiser les communautés marginalisées.

L'IA peut, par exemple, aider les organes de presse à adapter le contenu des informations aux préférences des utilisateurs, à analyser les données et à automatiser les tâches courantes, libérant ainsi un temps précieux pour un travail d'enquête plus approfondi. De plus, elle peut fournir des outils avantageux pour l'engagement politique et la représentation civile, en démocratisant l'accès et la participation. L'IA peut également être utilisée pour lutter contre la désinformation au moyen de techniques de détection et d'analyse automatisées. Enfin, les outils d'IA ont le potentiel de soutenir et d'améliorer les processus créatifs, ouvrant ainsi de nouvelles perspectives à l'expression artistique.

L'encouragement de telles utilisations positives par les gouvernements, les utilisateurs et les médias doit aller de pair avec la mise en place de garde-fous contre les pratiques d'utilisation irresponsables.

2.1. GOVERNMENTS

Les systèmes d'IA peuvent être utilisés à grande échelle pour propager des éléments de désinformation et manipuler l'opinion publique, ce qui peut s'avérer particulièrement préjudiciable en périodes électorales, de tensions sociopolitiques accrues, de conflits ou de crises économiques. Si les menaces que représente l'IA ne sont pas encore totalement comprises, sa nocivité potentielle n'en altère pas moins la confiance du public dans les institutions et les processus démocratiques. Pour garantir l'intégrité des informations gouvernementales et renforcer la confiance du public dans le processus démocratique, les gouvernements devraient prendre des mesures proactives pour encourager l'utilisation responsable des systèmes d'IA pour la création, la communication et la diffusion de contenu.

RECOMMANDATIONS AUX **ÉTATS**

> **Poursuivre les efforts visant à établir une charte ou des lignes directrices internationales pour l'utilisation de l'IA dans la création, la communication et la diffusion du contenu gouvernemental. Dans un premier temps, les pays ou les régions pourront prendre l'initiative d'élaborer et d'adopter cette charte, qui devrait inclure :**
>
> - Des normes sur l'étiquetage et le filigrane des contenus créés et diffusés par l'IA ou assistés par l'IA.

♦ Un engagement en faveur du contrôle humain et des mesures visant à réduire les risques de discrimination ou de biais.

> **Organiser des modules de formation pour les fonctionnaires afin de les familiariser avec les possibilités techniques des systèmes liés à l'IA, y compris leurs capacités et leurs limites, telles que les hallucinations et les infractions potentielles aux droits d'auteur.**

> **Adopter des normes d'authenticité et de provenance du contenu dans toutes les communications gouvernementales afin de renforcer la confiance et l'intégrité de l'information. En ce qui concerne la communication gouvernementale, ces normes devraient inclure l'authentification de l'auteur.**

2.2. ART ET SATIRE

Pour les artistes et les satiristes, l'IA offre des moyens efficaces et facilement accessibles de créer du contenu artistique ou satirique. Cependant, ces créations peuvent paraître authentiques, semant ainsi la confusion et risquant de porter tort au grand public. Rien de plus simple en effet que de fabriquer de la mésinformation et de la désinformation, des deepfakes et de produire des résultats biaisés et discriminatoires. Cette facilité exige un examen attentif de la manière dont l'IA générative, en particulier, est utilisée dans la création de contenus artistiques et satiriques, ainsi que du type d'étiquetage ou de provenance et d'authentification du contenu qui devrait être appliqué.[248]

RECOMMANDATIONS
À LA COMMUNAUTÉ ARTISTIQUE ET AUX SATIRISTES

> **Poursuivre les efforts visant à établir une charte ou des lignes directrices internationales pour l'utilisation responsable de l'IA pour la création et la diffusion de contenus artistiques et satiriques, en définissant des normes sur l'authenticité et la provenance des contenus, l'étiquetage, le filigrane et les mesures visant à garantir l'intégrité de l'information. Dans un premier temps, une telle charte pourrait être élaborée et adoptée au niveau national ou régional, ou par des associations représentatives du secteur.**

2.3. MÉDIAS ET JOURNALISTES

L'utilisation de l'IA pour la création de contenus synthétiques (textes, images, vidéos ou sons) par les journalistes recèle un potentiel considérable. L'IA peut améliorer la narration, faciliter la visualisation des données et simuler des scénarios à des fins d'investigation, enrichissant ainsi la production journalistique.

Le public ayant de moins en moins confiance dans les médias, l'utilisation de contenus synthétiques générés par l'IA pour l'influencer sans notification appropriée préalable pourrait éroder encore plus cette confiance, indispensable pour assurer le rôle de gendarme des organismes de médias et pour leurs modèles économiques.

248 La proposition de règlement de l'UE sur l'IA adopte une approche similaire aux exigences de transparence différenciées pour les œuvres ou programmes manifestement artistiques, créatifs, satiriques ou analogues à une fiction (article 52.3).

L'utilisation de l'IA à des fins de création et de diffusion d'informations doit donc se conformer aux normes éthiques du journalisme.

RECOMMANDATIONS
AUX ORGANES DE PRESSE ET AUX JOURNALISTES

> **Adopter un ensemble de lignes directrices et de normes professionnelles éthiques pour l'utilisation de l'IA au sein de chaque organisation médiatique ou dans l'ensemble du secteur.** Ces lignes directrices peuvent s'inspirer de la Charte de Paris sur l'IA et le journalisme[249] initiée par Reporters sans frontières (RSF) et 16 autres organisations. Les dix principes de la charte exigent, entre autres, que l'éthique journalistique régisse systématiquement les choix technologiques des médias, que les organes d'information privilégient le choix humain et assument leurs responsabilités, et que la personnalisation et la recommandation de contenu par l'IA favorisent l'intégrité et la diversité de l'information. En outre, la charte précise que les médias doivent établir une distinction claire entre le contenu synthétique et le contenu authentique et appliquer les normes les plus récentes en matière de traçabilité et d'authenticité. En ce qui concerne ce dernier principe, le document Responsible Practice for Synthetic Media[250] du Partnership on AI contient des conseils tactiques et techniques particulièrement utiles. Les lignes directrices du Conseil de l'Europe sur la mise en œuvre responsable des systèmes d'IA dans le domaine du journalisme fournissent également des orientations sur l'utilisation de l'IA par les médias et les journalistes.[251]

> **Définir les meilleures pratiques concernant l'utilisation de l'IA dans la création et la diffusion de contenus afin de fournir des orientations sur des questions et des cas spécifiques. Il s'agirait notamment :**

- **Des meilleures pratiques sur la création de contenus synthétiques (photo) réalistes.** Celles-ci devraient notamment préciser dans quels cas ce contenu doit être utilisé et comment il doit être étiqueté.
- **Des meilleures pratiques en matière de divulgation de l'implication de l'IA dans la création de contenu,** y compris lorsqu'il est nécessaire de divulguer l'utilisation d'outils d'IA (p.ex. l'utilisation de l'IA pour résumer un article, rédiger un titre, mener des recherches, analyser des données, etc.).
- **Des meilleures pratiques en matière de divulgation de l'implication de l'IA dans la diffusion de contenu, les systèmes de référencement des articles et les notifications.**

> **Mettre en place des modules de formation pour informer les journalistes sur les atouts et les limites des systèmes d'IA.** Une attention particulière devrait être accordée aux limites susceptibles de compromettre le travail journalistique, telles que les hallucinations de l'IA, les biais, les enjeux de la protection des sources et les droits d'auteur sur le contenu médiatique (voir section 3.2).

249 RSF (2023), RSF et 16 organisations partenaires présentent la Charte de Paris sur l'IA et le journalisme. Disponible sur : https://rsf.org/fr/rsf-et-16-organisations-partenaires-pr%C3%A9sentent-la-charte-de-paris-sur-l-ia-et-le-journalisme (Consulté le 8 février 2024).
250 Partnership on AI (n.d.). PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action. Disponible sur : https://syntheticmedia.partnershiponai.org/ (Consulté le 9 février 2024).
251 Conseil de l'Europe (2023). Lignes directrices sur la mise en œuvre responsable de systèmes d'intelligence artificielle dans le journalisme. Disponible sur : https://rm.coe.int/cdmsi-2023-014-lignes-directrices-sur-la-mise-en-uvre-responsable-de-s/1680adb4c7 (Consulté le 7 février 2024).

3. DÉVELOPPER LA CONNAISSANCE DE L'IA

Plus l'IA est présente dans notre vie quotidienne, plus elle influence notre façon d'interagir avec l'information, de débattre et même d'exercer nos droits fondamentaux, plus la question de l'éducation universelle à l'IA devient essentielle. Elle ne se limite pas aux professionnels qui travaillent directement avec des outils d'IA, dont les journalistes qui produisent et diffusent des informations, mais s'étend au grand public dans son ensemble.

Les programmes d'éducation à l'IA sont indispensables, tant pour les utilisateurs de l'IA que pour ses sujets, afin de mieux comprendre le fonctionnement des systèmes d'IA, démystifier les algorithmes et sensibiliser aux biais qui influencent leurs résultats. Les utilisateurs doivent notamment être conscients du fait qu'il est possible d'utiliser l'IA pour créer des deepfakes et des éléments de mésinformation et de désinformation, de les diffuser à grande échelle et de générer l'approbation et le partage de ces contenus par l'IA elle-même. Cette prise de conscience permet à la fois d'atténuer les préjudices potentiels et de favoriser les effets bénéfiques. De plus, une bonne compréhension des aspects juridiques et éthiques de l'IA permet aux individus de demander des comptes aux développeurs et déployeurs d'IA, de revendiquer un paysage de l'information plus équitable et plus juste, et d'encourager une utilisation responsable de l'IA par tous. Les responsables des pouvoirs publics doivent eux aussi posséder une bonne compréhension de ces outils, afin de pouvoir remplir efficacement leurs obligations vis-à-vis des citoyens. C'est pourquoi des programmes ciblés d'éducationà l'IA sont essentiels pour préserver les droits et défendre la démocratie à l'ère du numérique.

Une éducation ciblée à l'IA destinée à tous peut permettre à chacun de faire preuve d'esprit critique à l'égard de l'information, de contribuer activement à l'univers numérique et de défendre les principes de responsabilité, de diversité et d'accès à des informations fiables. Cet effort collaboratif garantira que chacun dispose des connaissances nécessaires pour appréhender les complexités des systèmes d'IA et bâtir un avenir où ceux-ci seront des outils au service de l'engagement démocratique.

3.1. L'ÉDUCATION À L'IA ET LE RÔLE DES ENTREPRISES

Les entreprises et entités spécialisées dans l'IA ont une grande responsabilité concernant la promotion et le financement de l'éducation à l'IA.[252] Elles doivent notamment fournir des informations facilement accessibles et compréhensibles sur le fonctionnement, les limites et les risques potentiels de leurs systèmes. Les utilisateurs ont besoin d'explications claires sur le fonctionnement de ces systèmes, les données qu'ils utilisent pour apprendre, la manière dont les prompts sont utilisés et les biais éventuels que ces systèmes peuvent présenter.

En misant sur la transparence et en investissant dans des outils facilement accessibles pour mieux comprendre le fonctionnement des systèmes d'IA, les entreprises permettent aux utilisateurs d'interagir avec l'IA de manière responsable..

252 La proposition de règlement de l'UE sur l'IA oblige les fournisseurs et déployeurs d'IA à améliorer les compétences en matière d'IA de leur personnel (article 4b).

RECOMMANDATIONS
AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Expliquer de manière exhaustive et accessible les fonctionnalités des systèmes d'IA, leurs limites, l'utilisation des données et les risques potentiels. Proposer à cette fin de multiples supports, tels que des infographies et des vidéos, afin de répondre aux besoins de différents publics.**

> **Proposer des modules de formation et des programmes éducatifs destinés à différents groupes d'utilisateurs, y compris les responsables politiques et le grand public.**

3.2. PROGRAMMES D'ÉDUCATION À L'IA DESTINÉS AU GRAND PUBLIC, AUX POUVOIRS PUBLICS, AUX MÉDIAS, À L'APPAREIL JUDICIAIRE ET À D'AUTRES ACTEURS

D'après une étude de l'UNESCO réalisée en 2022 et à laquelle 51 pays membres ont répondu, seuls 11 d'entre eux ont rédigé et mis en œuvre un programme d'éducation à l'IA dans le domaine de l'éducation.[253] Introduire un programme d'éducation à l'IA dans les programmes scolaires peut donc être considéré comme une étape importante afin de mieux faire connaître l'IA aux jeunes générations.

Au-delà de l'éducation à l'IA au sein du système éducatif, les programmes de formation destinés au grand public sont essentiels pour favoriser l'utilisation éthique des systèmes d'IA et réduire les risques qu'ils font peser sur l'intégrité de l'information et la confiance dans les institutions démocratiques. Les gouvernements pourraient s'inspirer du programme finlandais IA « 1 % », qui vise à former 1 % de la population dans le cadre d'un cours universitaire gratuit sur les principes de base des systèmes d'IA.[254] Les gouvernements, en collaboration avec les universités et les organisations de la société civile, devraient concevoir des cours spécialisés sur le fonctionnement, les possibilités et les limites des systèmes d'IA, en accordant une attention particulière aux implications pour l'espace informationnel. La collaboration avec des entreprises privées et la formation continue pourraient contribuer à améliorer cette assimilation. Une attention particulière doit être accordée à la maîtrise de l'IA par les groupes vulnérables et marginalisés, pour qui une collaboration avec des organisations locales s'avère particulièrement judicieuse. Le FactShala de DataLEADS est un exemple de ce type d'initiative : il s'agit d'un programme de maîtrise des médias et de l'information qui aide les habitants des petites villes et des villages de l'Inde à analyser les informations en ligne et à distinguer les faits de la désinformation avec un esprit critique.[255] Des secteurs spécifiques peuvent avoir besoin d'une éducation ciblée à la maîtrise de l'IA, comme le cours de gestion globale de l'infodémie de DataLEADS pour les professionnels de la santé,[256] qui traite du défi spécifique de la mésinformation générée par l'IA dans le secteur des soins de santé.

253 UNESCO (2022). K-12 AI curricula: A mapping of government-endorsed AI curricula. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380602 (Consulté le 8 février 2024).
254 Delcker, J. (2019), Finland's grand AI experiment, Politico. Disponible sur : www.politico.eu/article/finland-one-percent-ai-artificial-intelligence-courses-learning-training/ (Consulté le 9 février 2024).
255 DataLeads. FactShala: India's Largest Media Literacy Network. Disponible sur : https://dataleads.co.in/capacity-building/#FactShala (Consulté le 9 février 2024).
256 Global Infodemic Management Course (n.d.). Global Infodemic Management Course for Healthcare Workers. Disponible sur : https://gimch.org/ (Consulté le 9 février 2024).

RECOMMANDATIONS AUX **ÉTATS**

> **Investir dans des programmes d'éducation à l'IA pour les journalistes en partenariat avec des écoles de journalisme.**

> **Intégrer des modules d'éducation à l'IA dans les programmes de formation existants destinés aux fonctionnaires, y compris les magistrats,** afin de leur fournir en continu des opportunités de développement des compétences et d'acquisition de connaissances.

> **Investir dans l'IA et la culture numérique des citoyens, en particulier des groupes vulnérables et des minorités, afin qu'ils disposent des compétences techniques et des ressources cognitives nécessaires pour mieux se repérer dans les espaces d'information, identifier la mésinformation et la désinformation de manière indépendante, et utiliser les systèmes d'IA de manière responsable.** Il peut s'agir de :

- **Intégrer la formation à l'IA dans les programmes d'enseignement à tous les niveaux d'éducation** (primaire, collège, lycée et enseignement supérieur).
- **Lancer des campagnes de sensibilisation du public relatives aux risques liés aux systèmes d'IA** (p.ex. les biais, la manipulation par l'IA et la mésinformation et la désinformation) **ainsi que sur le potentiel de l'IA à démocratiser l'information** (p.ex. en générant des informations ciblées qui répondent aux préoccupations spécifiques des utilisateurs, en particulier ceux issus de communautés défavorisées).
- **Organiser des formations sur l'IA et ses implications pour l'espace informationnel,** disponibles gratuitement et à grande échelle, aux niveaux mondial ou national. Ces cours devraient :
  - Être accessibles via une plateforme interactive en ligne ou une application centralisée proposant plusieurs cours centrés sur les résultats et visant à développer les compétences professionnelles et la compréhension éthique de l'IA au sein de différents groupes d'âge.[257]
  - Être disponibles dans plusieurs langues et prendre en compte les différences culturelles des différents groupes de la société.

> **Favoriser les programmes ciblés destinés aux groupes vulnérables en préparant du matériel de formation,[258] et financer des programmes répondant aux besoins de catégories spécifiques d'utilisateurs.[259]**

> **Travailler en partenariat avec les syndicats et les associations professionnelles pour intégrer la maîtrise de l'IA dans la formation professionnelle dans différents secteurs.**

257 Similaire à la plateforme d'apprentissage en ligne gratuite AI Campus financée par le Ministère fédéral allemand de l'Éducation et de la Recherche. Disponible sur : https://ki-campus.org/publications (Consulté le : 7 février 2024).

258 Découvrez FactShala de DataLEADS, un programme de littératie médiatique et informationnelle dirigé par DataLEADS avec le soutien de l'Initiative Google News, qui aide les personnes des villes et villages à travers l'Inde à évaluer de manière critique les informations en ligne et à distinguer les faits de la désinformation. Plus d'informations disponibles sur : https://dataleads.co.in/capacity-building/#FactShala (Consulté le : 7 février 2024).

259 Découvrez le cours de gestion mondiale de l'infodémie de DataLEADS pour les travailleurs de la santé, qui aborde le défi spécifique de la désinformation alimentée par l'IA dans le secteur de la santé. Plus d'informations disponibles sur : https://gimch.org/ (Consulté le : 7 février 2024).

> **Veiller à ce qu'un programme efficace d'éducation à l'IA couvre de manière exhaustive divers aspects clés de l'IA, y compris, mais sans s'y limiter, des sujets tels que :**
>
> - Les fondamentaux de l'IA (définition, fonctionnalités, applications courantes, chaîne de valeur de l'IA)
> - Les malentendus généraux sur l'IA (neutralité de la valeur de l'IA, risques d'erreurs de l'IA, concept d'hallucinations de l'IA)
> - Les biais et la discrimination dans la création de contenu facilitée par l'IA
> - Les préoccupations en matière de protection de la vie privée, de protection des données et de droits d'auteur associées à l'IA
> - L'utilisation de l'IA dans la propagation de contenus préjudiciables, la diffamation, les discours de haine et les prises de position illégales
> - L'utilisation de l'IA dans la création de mésinformation et de désinformation, pour la prolifération de bots sur les réseaux sociaux et le micro-ciblage
> - Les outils et les compétences pour détecter les deepfakes et d'autres formes de contenu synthétique
> - La capacité à comprendre les systèmes de recommandation et de modération de contenu de l'IA
> - Les régimes de responsabilisation et de responsabilité (établir la responsabilité des dommages et de la non-conformité des systèmes d'IA)

3.3. FINANCEMENT DE L'ÉDUCATION À L'IA

L'accès équitable à la maîtrise de l'IA nécessite une approche plurielle impliquant un financement public, une coopération internationale et une participation active du secteur privé, notamment par la contribution d'une part significative de ses bénéfices à cette priorité politique. Cela implique d'investir dans les infrastructures publiques, d'établir de solides partenariats transfrontaliers et de soutenir les projets d'éducation à l'IA, y compris dans les pays de « la majorité mondiale ». Il est également important de tirer parti des ressources actuelles, en accordant un accès libre et ouvert à l'information et en soutenant les initiatives locales.

Bien que l'intelligence artificielle représente un formidable potentiel d'amélioration de la société, comme cela a été évoqué au fil du présent rapport, les systèmes d'IA ont également des externalités négatives, notamment le fait de laisser les mains libres aux acteurs malveillants. À l'instar des taxes environnementales, les entreprises et entités du secteur de l'IA devraient supporter le coût de ces externalités au bénéfice de la société. Une partie des recettes perçues par le biais d'une taxe sur les entreprises d'IA pourrait contribuer à un fonds éthique pour l'IA, semblable aux fonds de service et d'accès universels auxquels les opérateurs de téléphonie mobile et les fournisseurs de télécommunications contribuent pour améliorer l'accès à Internet dans le monde entier.[260]

260 Web Foundation et al; (2018). Universal Service and Access Funds: An Untapped Resource to Close the Gender Digital Divide. Disponible sur : https://webfoundation.org/docs/2018/03/Using-USAFs-to-Close-the-Gender-Digital-Divide-in-Africa.pdf (Consulté le 9 février 2024)

RECOMMANDATIONS AUX **ÉTATS**

> **Affecter des fonds supplémentaires aux pays de « la majorité mondiale » via une aide publique au développement (APD), en les allouant spécifiquement aux programmes d'éducation à l'IA.**

> **Instaurer une taxe sur les entreprises et les entités actives dans le domaine de l'IA afin de tenir compte des incidences sociétales de l'IA.** Une partie des recettes générées par cette taxe devrait être affectée au financement de programmes communautaires d'éducation à l'IA, d'alternatives publiques aux systèmes à but lucratif et d'initiatives de la société civile (voir section 1.5 et chapitre 4, section 3.2). Ces initiatives visent à créer une société mieux informée et plus autonome, tout en mettant en place des incitations économiques pour que les entreprises et les entités spécialisées dans l'IA réduisent au minimum les effets négatifs de leurs systèmes d'IA.

Pour garantir l'efficacité de cette taxe, celle-ci doit :

- **S'appuyer sur les travaux en cours en collaboration avec l'OCDE et les renforcer afin d'établir un taux d'imposition minimum effectif de 15 % pour les multinationales de l'IA dans chaque juridiction, quel que soit le lieu où elles opèrent.**[261]
- **S'appliquer aux entreprises et aux entités qui déploient des systèmes d'IA, avec des exemptions ciblées pour la recherche, l'éducation et d'autres applications au service de l'intérêt public.**
- **Se fonder sur le nombre d'utilisateurs d'un système d'IA, sur les revenus qui y sont associés et sur sa classification des risques.** Cette approche imposerait une charge financière plus lourde aux entreprises et aux entités qui déploient des systèmes d'IA plus importants, susceptibles de causer davantage de dommages.
- **Être assortie de sanctions en cas de non-respect, y compris des amendes et la suspension des déploiements d'IA.**
- **Être conçue et mise en œuvre en coordination avec les parties prenantes concernées par l'IA, y compris les représentants de la société civile,** afin de garantir un mandat solide.

261 OCDE (2023). Outcome Statement on the Two-Pillar Solution to Address the Tax Challenges Arising from the Digitalisation of the Economy. Disponible sur : www.oecd.org/tax/beps/outcome-statement-on-the-two-pillar-solution-to-address-the-tax-challenges-arising-from-the-digitalisation-of-the-economy-july-2023.pdf (Consulté le 9 février 2024).

CHAPITRE 4 : GOUVERNANCE ET CONTRÔLE DE L'IA

INTRODUCTION

La rapidité et l'ampleur avec lesquelles les systèmes d'IA progressent sont sans précédent. Les opportunités et les risques inédits qu'ils présentent pour l'intégrité des écosystèmes de l'information et de la communication se traduisent par de nouveaux défis et enjeux pour les responsables politiques, qui doivent non seulement réagir à l'évolution constante du paysage de l'IA, mais également anticiper ses implications futures.

Les systèmes d'IA sont de plus en plus aptes à influencer le débat sur les politiques publiques, ainsi que l'opinion publique sur diverses thématiques. Le développement fulgurant de l'IA, associé à son adoption rapide, complique considérablement la tâche des responsables politiques. De plus, le manque de connaissances sur l'IA, de ressources financières et humaines des autorités et le lobbying des entreprises du secteur aggravent encore la situation. Ces intérêts privés, bien qu'ils défendent publiquement une approche réglementaire pilotée par l'État, tiennent également un discours suivant lequel la réglementation étouffe l'innovation, et que cette innovation doit impérativement évoluer comme ils l'envisagent. En outre, ces intérêts détournent l'attention des défis concrets en présentant les risques liés à l'IA comme une question d'éthique plutôt que de droit. Enfin, l'accès restreint aux systèmes d'IA propriétaires entrave les travaux de recherche d'intérêt public, ce qui complique davantage les initiatives menées par les responsables politiques pour comprendre les systèmes d'IA et s'attaquer aux risques qu'ils présentent.

Si la plupart des pays disposent de réglementations qui pourraient relever les défis posés par l'IA, leur mise en œuvre s'accompagne souvent d'importantes ambiguïtés. Ainsi, la question de l'attribution des droits d'auteur sur les contenus générés par l'IA fait actuellement débat, en particulier lorsque ces contenus reposent sur des droits de propriété intellectuelle préexistants. De plus, les systèmes d'IA présentent un ensemble de risques uniques auxquels les autorités doivent répondre par des réglementations spécifiques, notamment en ce qui concerne l'attribution de la responsabilité du contenu généré par l'IA et des décisions qui en découlent. Cela suppose que l'approche adoptée inscrive clairement dans la loi les droits à protéger, notamment la liberté d'expression, le respect de la vie privée, la protection des données et la non-discrimination. Il s'agit également d'établir des règles claires concernant les comportements autorisés et interdits, dont le respect est assuré par des régimes de responsabilité.

Dans le cadre de la réglementation des systèmes d'IA, et dans le but de limiter les préjudices existants et potentiels pour les personnes, la société et les institutions démocratiques, les responsables politiques doivent adopter une approche équilibrée, en adaptant la portée de la réglementation et des obligations aux risques posés par les systèmes d'IA. Des mesures trop rigides pourraient en effet avoir une incidence sur l'innovation et entraver l'entrée sur le marché de petites entreprises et de start-ups, renforçant ainsi la domination de grandes entreprises et entités spécialisées dans l'IA. Par conséquent, les mesures les plus strictes devraient être réservées aux systèmes d'IA susceptibles d'avoir les répercussions les plus graves sur l'espace de l'information. En parallèle, un excès de prudence pourrait conduire les gouvernements à trop se fier à la bonne volonté des entreprises et à négliger de fixer des règles démocratiques claires en termes de développement, de déploiement et d'utilisation des systèmes d'IA. Les États doivent donc établir des cadres réglementaires et de gouvernance visant à garantir des espaces d'information et de communication fiables, sûrs, équitables et démocratiques, tout en adoptant une approche innovante du développement de l'industrie de l'IA. À cette fin, il convient d'adopter une approche articulée sur des principes, qui définit un ensemble complet de valeurs et d'objectifs devant guider tout cadre institutionnel chargé

de protéger les écosystèmes informationnels contre les effets potentiellement préjudiciables des systèmes d'IA. Cette approche se fonde sur des principes généraux, plutôt que sur des spécificités techniques et contextuelles concernant la manière dont les autorités administratives locales devraient être structurées.

Premièrement, la défense de l'intérêt public, du droit international relatif aux droits humains et des valeurs démocratiques doit être un objectif central pour les responsables politiques. L'élaboration des politiques doit notamment prendre en compte les besoins des populations dans toute leur diversité, en tenant compte des implications sociopolitiques de l'IA, en particulier pour les groupes vulnérables. Les responsables politiques devraient également mettre en place des procédures efficaces pour favoriser la participation des organisations de la société civile, des chercheurs, des journalistes et d'autres groupes marginalisés à l'élaboration, à la mise en œuvre et à la surveillance des politiques.

Deuxièmement, les responsables politiques doivent se garder d'adhérer à l'hypothèse philosophique répandue, mais fallacieuse, selon laquelle les technologies peuvent être apolitiques et neutres sur le plan des valeurs. Non seulement les technologies véhiculent les valeurs de leurs concepteurs, de leurs déployeurs et de leurs utilisateurs, mais leurs implications sociopolitiques directes et indirectes transmettent également certaines valeurs.[262]

Troisièmement, il appartient aux responsables politiques de remettre en question l'idée prédominante selon laquelle la réglementation doit intervenir une fois que les innovations technologiques ont fait leur apparition.[263] *Au contraire, les dirigeants devraient essayer d'être aussi créatifs que les innovateurs technologiques, en cartographiant et en répondant de manière proactive aux risques et aux opportunités présentés par les systèmes d'IA afin d'orienter l'innovation dans une direction responsable qui favorise l'intérêt public et renforce les institutions démocratiques.*

Ces recommandations souscrivent au principe de « l'équivalence fonctionnelle » inscrit dans les réglementations internationales. Ce principe reconnaît que des approches différentes peuvent aboutir aux mêmes résultats dans le cadre d'une gouvernance démocratique. En permettant une certaine flexibilité, l'équivalence fonctionnelle garantit véritablement le respect des divers contextes dans lesquels l'IA sera développée, déployée et utilisée.

Pour surmonter efficacement les difficultés exposées ci-dessus, les États doivent mettre en place une gouvernance efficace de l'IA, assortie d'exigences adéquates en matière de transparence et de contrôle. Ces exigences figurent notamment dans le Projet de convention-cadre du Conseil de l'Europe sur l'intelligence artificielle, les droits de l'homme, la démocratie et l'État de droit.[264] *Les exigences en matière de transparence et de contrôle devraient comprendre plusieurs éléments, parmi lesquels :*

- *Des institutions démocratiques fortes chargées de superviser, d'élaborer et d'appliquer les réglementations et les régimes de responsabilité.*
- *Un contrôle démocratique, une participation citoyenne et une implication équitable, durable et substantielle de la société civile.*
- *Des évaluations de conformité garantissant le respect des exigences légales avant le déploiement.*
- *Des mécanismes d'audit du comportement des systèmes d'IA et de leurs processus de développement exigeants.*

---

262 Magrani, E. (2019). New perspectives on ethics and the laws of artificial intelligence. Internet Policy Review. Disponible sur : https://policyreview.info/articles/analysis/new-perspectives-ethics-and-laws-artificial-intelligence (Consulté le 7 février 2024).

263 Kretschmer, M. et al (2023). *The risks of risk-based AI regulation: taking liability seriously.* Social Science Research Network. Disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4622405 (Consulté le 7 février 2024).

264 Conseil de l'Europe (2023). Projet de Convention-cadre sur l'Intelligence Artificielle, les Droits de l'Homme, la Démocratie et l'Etat de Droit. Article 7. Disponible sur : https://rm.coe.int/cai-2023-28-fr-projet-de-convention-cadre/1680ae19a1 (Consulté le 7 février 2024).

- *Un ensemble d'outils et de dispositifs permettant aux utilisateurs de porter plainte et d'obtenir réparation.*
- *Des investissements dans la recherche et l'accès aux systèmes d'IA pour permettre aux chercheurs d'étudier les systèmes et de leur faire assumer leurs responsabilités.*

Afin de concrétiser pleinement leur potentiel, ces composantes devraient s'inscrire dans un cadre de gouvernance international, garantissant ainsi une approche globale et coordonnée de la gouvernance de l'IA.

1. MISE EN PLACE D'INSTITUTIONS DÉMOCRATIQUES FORTES

Face à l'ampleur des effets néfastes que l'IA peut avoir sur nos démocraties, les institutions doivent être fortes et résilientes, qu'elles soient existantes ou nouvelles, et aptes à relever les défis actuels. Ces institutions doivent être capables de faire adopter efficacement de nouvelles lois, de faire appliquer les lois existantes et d'assurer un encadrement rigoureux des systèmes d'IA, de leurs développeurs, de leurs déployeurs et de leurs utilisateurs.

En créant des institutions nationales fortes et indépendantes, expertes en IA et aux responsabilités clairement définies, les gouvernements contribuent à créer un espace informationnel sain et fiable, renforcent la confiance du public dans l'écosystème informationnel et œuvrent pour la protection des principes démocratiques dans l'espace de l'information et de la communication.

1.1. LES DÉFIS POSÉS PAR L'IA EN TERMES DE RÉGLEMENTATION

Le développement rapide des systèmes d'IA a mis en lumière les nombreuses incertitudes juridiques et zones d'ombre des réglementations existantes pouvant être facilement contournées ou détournées par des intérêts privés motivés par des intérêts commerciaux, malveillants ou autres. Ces incertitudes portent aussi sur des questions fondamentales telles que la protection des données, la vie privée, la propriété intellectuelle, les droits d'auteur et les dispositions relatives à la responsabilité civile. En outre, de nombreuses questions restent sans réponse concernant les réglementations déjà adoptées (p.ex. le règlement sur les services numériques (DSA) et la loi britannique sur la sécurité en ligne) ou en cours d'élaboration ou de finalisation (p.ex. la loi européenne sur l'IA ou le projet de législation brésilienne sur l'IA).

Soucieux de protéger les droits humains et un espace informationnel démocratique, les États doivent adopter une approche proactive afin de créer une sécurité juridique en matière d'IA. Cela suppose de clarifier l'applicabilité de la législation existante aux systèmes d'IA par une interprétation cohérente des lois, et de promulguer de nouvelles lois et réglementations, le cas échéant. Ainsi, selon la CNUCED[265], bien que 137 pays au monde disposent d'une loi sur la protection des données, celle-ci n'est pas toujours adaptée aux meilleures pratiques internationales, ce qui la rend potentiellement inefficace pour relever les défis liés à l'IA.

Dans ce contexte, le respect des meilleures pratiques internationales, même s'il n'est pas toujours suffisant pour répondre aux questions émergentes en matière d'IA, peut garantir le respect fondamental des droits humains et du droit international. De plus, la mise en œuvre de lignes directrices destinées à compléter les réglementations existantes est essentielle pour garantir leur application cohérente et uniforme à l'IA.

265 Conférence des Nations Unies sur le commerce et le développement. Data Protection and Privacy Legislation Worldwide. Disponible sur : https://unctad.org/page/data-protection-and-privacy-legislation-worldwide (Consulté le 7 février 2024).

RECOMMANDATIONS AUX **ÉTATS**

> **Adopter une législation appropriée, fondée sur les lignes directrices et les normes internationales lorsqu'elles existent, afin de répondre spécifiquement aux enjeux posés par l'IA dans des domaines juridiques majeurs** tels que la propriété intellectuelle, les droits d'auteur, la protection des données, la vie privée, la non-discrimination, la gouvernance des plateformes et la responsabilité civile.

> **Examiner les lois et règlements existants et évaluer leur applicabilité aux risques posés par les systèmes d'IA dans l'espace informationnel, afin d'identifier les lacunes et les ambiguïtés.** Ces lois concernent, entre autres, la protection des données, la vie privée, la propriété intellectuelle, les droits d'auteur, la gouvernance des plateformes numériques et la responsabilité civile.

> **Définir des lignes directrices exhaustives en termes d'interprétation et de mise à jour des réglementations existantes afin de répondre efficacement aux questions liées à l'IA dans l'espace de l'information et de la communication. Ces lignes directrices devraient à minima :**

- **Établir des critères clairs concernant les droits d'auteur des produits générés par l'IA,** en particulier lorsqu'ils reposent sur des droits de propriété intellectuelle préexistants.
- **Préciser le champ d'application des exceptions relatives à l'usage raisonnable des pratiques de curation des données associées à l'apprentissage des systèmes d'IA,** y compris la manière dont ces règles s'appliquent aux systèmes d'IA générative et à leurs résultats.
- **Préciser les droits et la propriété relatifs aux métadonnées générées par les systèmes d'IA.**
- **Clarifier l'application des principes de protection des données tels que le droit à l'oubli, le consentement, les mécanismes d'exclusion** et d'autres droits relatifs au contenu et aux systèmes générés par l'IA, afin de permettre aux sujets de l'IA de maîtriser leur empreinte numérique (voir chapitre 1, section 1.e).
- **Identifier les lacunes et proposer des modifications aux lois, règles et réglementations électorales** afin de tenir compte des risques et des implications liés à l'utilisation de contenus générés par l'IA dans le cadre des campagnes électorales (voir chapitre 2, section 4.2).
- **Établir la responsabilité de la modération, de la vérification, de la curation, de la recommandation et des décisions de ciblage et de diffusion des publicités des systèmes d'IA** (voir chapitre 2, section 3.1).
- **Clarifier la responsabilité des résultats des systèmes d'IA** (voir chapitre 2, section 3.1).
- **Déterminer le degré d'autonomie des systèmes d'IA, en particulier dans les situations à forts enjeux, et préciser le rôle de la supervision humaine** (voir chapitre 2, section 4).
- **Fixer des cadres pour garantir la qualité et la conformité constantes des systèmes d'IA par l'évaluation de la conformité et l'audit** (voir section 2). Ceci inclut la nécessité de mener des recherches de fond sur les méthodologies d'évaluation visant à évaluer les performances des systèmes d'IA générative.

1.2. AUTORITÉS DE RÉGULATION ET DE CONTRÔLE

Les réglementations ne seront efficaces que si les autorités de régulation, de contrôle et d'administration disposent d'une réelle indépendance institutionnelle, de la capacité d'agir concrètement en fonction de leurs attributions, et de ressources suffisantes pour assurer leur application et leur évolution.

C'est la raison pour laquelle les États sont invités à renforcer les institutions existantes ou à en créer de nouvelles spécifiquement adaptées à l'IA, afin que les gouvernements conservent leur expertise dans les domaines liés à l'IA, tels que la protection des données et les droits humains. Ces organismes doivent impérativement être dotés de nouveaux pouvoirs leur permettant non seulement de superviser l'application des réglementations en matière d'IA, mais aussi de faire adopter des actes d'exécution.

En fonction du contexte national, ces pouvoirs peuvent être répartis entre différentes autorités compétentes pour un secteur donné, ou confiés à une seule autorité. Dans tous les cas, ces autorités doivent bénéficier d'une indépendance institutionnelle garantie par la loi, afin que leurs actions soient indépendantes et libres de toute ingérence politique de la part du gouvernement, des élus, des entreprises et d'autres entités.[266] De plus, ces autorités devraient se voir attribuer des pouvoirs juridiques leur permettant de mettre en œuvre et d'appliquer efficacement les réglementations. Elles devraient notamment disposer de pouvoirs d'investigation, tels que la possibilité de solliciter les informations et la documentation nécessaires sur les systèmes d'IA et de mener des études techniques sur le fonctionnement de ces systèmes. Outre son indépendance, la structure institutionnelle choisie devrait également garantir l'expertise, des ressources suffisantes, l'adaptabilité, la transparence, la responsabilité, les capacités d'anticipation et la proactivité de ces autorités.

RECOMMANDATIONS AUX **ÉTATS**

> **Établir une nouvelle autorité ou renforcer les capacités des autorités existantes** (p.ex. les autorités chargées de la protection des données, les commissions des droits humains ou les autorités chargées de la concurrence) **et les charger de superviser l'application des réglementations en matière d'IA et d'adopter des lois et des règlements d'application.**[267] Cette autorité devrait être responsable du suivi des tendances, de l'identification des risques évolutifs ou futurs et de la formation proactive des responsables de l'application des lois et des fonctionnaires. Elle devrait également :
>
> - **Définir un cadre réglementaire exhaustif pour le développement, le déploiement et l'utilisation de l'IA, conformément à la législation primaire.**
> - **Superviser la mise en œuvre des politiques et stratégies nationales en matière d'IA.**
> - **Adopter et promulguer des lois de mise en œuvre et de réglementation.**
> - **Faire appliquer les réglementations existantes** avec le concours d'autres autorités de régulation.

266 Le projet de règlement sur l'IA de l'UE comprend diverses dispositions visant à assurer une application efficace de la régulation (Chapitre 3 : Application, Articles 63-68a).

267 Dans l'UE, un bureau de l'IA, faisant partie de la Commission européenne, sera chargé de surveiller la conformité avec le règlement sur l'IA de l'UE selon sa version préliminaire (Article 55b). Aussi :
Commission européenne (2024). Commission Decision Establishing the European AI Office. Disponible sur : https://digital-strategy.ec.europa.eu/en/library/commission-decision-establishing-european-ai-office (Consulté le : 7 février 2024).

- **Imposer des pénalités en cas de non-respect des lois, règles et réglementations pertinentes en matière d'IA,** pouvant inclure des poursuites pénales en cas de faute délibérée entraînant des atteintes aux droits humains, ainsi que l'interdiction pour les systèmes d'IA d'entrer ou de rester sur le marché.
- **Recevoir et traiter les plaintes,** à moins qu'un médiateur ne soit désigné à cet effet.
- Recevoir des rapports d'incidents et préconiser des mesures correctives.
- Tenir un registre public des systèmes d'IA existants et de leur documentation.
- **Délivrer et révoquer les licences** dans les juridictions où ce système est en place.
- **Appliquer les procédures nécessaires à l'évaluation, à la désignation et à la notification des organismes tiers d'évaluation de la conformité, ainsi qu'à leur contrôle.**
- **Définir et publier des critères** (à partir des critères élaborés dans le tableau 1.2) **de classement des systèmes d'IA en fonction des risques potentiels qu'ils présentent et de leur incidence systémique sur l'espace informationnel**.
- **Promouvoir l'éducation de l'IA auprès du grand public** en coopération avec les entreprises privées, le gouvernement et les partenaires du monde de l'éducation.
- **Apporter son expertise et sa contribution à d'autres institutions publiques** (gouvernement, pouvoir judiciaire) le cas échéant.
- **Favoriser la coopération avec les organisations internationales afin d'harmoniser les politiques et les normes en matière d'IA,** en privilégiant les travaux de mise en conformité détaillés qui devraient impliquer la création de normes IEEE et ISO et la référence à ces normes.

> **Imposer aux développeurs et déployeurs de systèmes d'IA à haut risque de contribuer au coût de gouvernance par le versement d'une redevance de supervision.**[268]

> **Mettre en place des réglementations et des garanties, et affecter des ressources pour s'assurer que la ou les autorités en question disposent des moyens indispensables pour fonctionner de manière indépendante et efficace, notamment :**

- **L'indépendance** – les autorités doivent fonctionner de manière autonome, libres de toute influence politique, commerciale ou autre influence extérieure susceptible de compromettre leur objectivité ou leur capacité à agir dans l'intérêt public.
- **L'expertise** – les autorités doivent être composées de personnes de tous horizons possédant une connaissance étendue des systèmes d'IA et de leurs implications sociétales, capables de faire appliquer la loi, de publier des actes d'exécution et de fournir des conseils et des avis. Afin de garantir une dotation en personnel appropriée de ces entités, il convient de :
  - Mettre en place de solides procédures de sélection et de vérification des antécédents pour le recrutement du personnel afin d'éviter tout conflit d'intérêts.
  - Engager des ressources pour s'assurer que le personnel dispose de l'expertise nécessaire et des outils de pointe pour accomplir ses missions.
- **Des ressources adéquates** – un financement approprié et un personnel qualifié sont essentiels pour permettre à ces organismes de réglementation d'exercer leurs fonctions de contrôle.

268 Le règlement de l'UE sur les services numériques oblige les plateformes en ligne à payer la redevance de supervision annuelle (Article 43).

- **Le pouvoir de mise en application** – les autorités doivent disposer du pouvoir juridique nécessaire pour mettre en œuvre et faire respecter les réglementations avec efficacité.
- **L'adaptabilité** – compte tenu du développement rapide des systèmes d'IA, les autorités doivent faire preuve d'agilité et être en mesure de s'adapter rapidement aux avancées technologiques pour veiller à ce que les réglementations restent pertinentes et efficaces. Elles doivent surveiller les tendances de manière proactive, identifier les risques évolutifs et futurs, et former les responsables de l'application des lois.
- **La transparence et la responsabilité** – les opérations et les processus décisionnels des organismes de régulation doivent être ouverts et transparents afin de gagner la confiance du public et de garantir la responsabilité. Cela suppose de communiquer de manière proactive avec les différentes parties prenantes, de publier des rapports sur les avancées des travaux et de solliciter l'avis du public.
- **Des pouvoirs d'investigation** – pour être efficaces, ces organismes devraient se voir octroyer des pouvoirs d'investigation étendus, parmi lesquels un accès total[269] aux algorithmes (et à leur documentation) des systèmes d'IA, ainsi qu'aux données à partir desquelles ils apprennent et opèrent.

1.3. PARTICIPATION MULTIPARTITE À LA GOUVERNANCE DE L'IA

Afin de garantir en permanence la prise en compte des différents points de vue et un contrôle démocratique, il est essentiel que les parties prenantes les plus diverses (notamment le public, la société civile, le milieu universitaire, les journalistes, les entreprises et les groupes de plaidoyer) soient représentées dans le processus de contrôle de l'IA. À cette fin, les États doivent créer des mécanismes efficaces permettant d'impliquer ces différentes parties prenantes de manière équitable, durable et substantielle. Il pourrait s'agir d'un comité consultatif spécifique, créé pour conseiller l'autorité chargée de l'IA, ou d'une institution rattachant directement les différentes parties prenantes à la structure de gouvernance de l'autorité.

RECOMMANDATIONS
AUX **ÉTATS**

> **Mettre en place un mécanisme garantissant une participation équitable, durable et substantielle des parties prenantes, incluant la société civile, les chercheurs, les communautés affectées et marginalisées, et les experts, au sein de l'organisme de réglementation de l'IA. Deux options** peuvent être envisagées pour atteindre cet objectif :
>
> - La création d'un **conseil consultatif indépendant**[270] chargé de conseiller l'autorité chargée de l'IA.

269 Le règlement de l'UE sur les services numériques oblige les plateformes en ligne à permettre aux auditeurs tiers d'accéder à toutes les données pertinentes nécessaires à la réalisation des audits (article 37.2).

270 La proposition de règlement de l'UE sur l'IA prévoit la création du Comité européen de l'intelligence artificielle, composé de représentants des États membres (article 56). En outre, les États membres devraient établir/désigner au moins une autorité notifiante et au moins une autorité de surveillance du marché aux fins de la mise en œuvre du règlement sur l'IA (article 59).

♦ Le **rattachement direct des parties prenantes à la structure de gouvernance de l'autorité en charge de l'IA.**[271]

> **Mettre en place un mécanisme de sélection transparent, inclusif et redevable pour les parties prenantes invitées à siéger au conseil consultatif ou à l'autorité chargée de l'IA.** Ce processus de sélection doit obéir à des critères clairement définis et viser une représentation diversifiée des groupes et des intérêts.

1.4 CAPACITÉS DE RECHERCHE

Les systèmes d'IA étant souvent opaques, il peut y avoir de grandes disparités en termes de compréhension entre leurs développeurs et leurs utilisateurs d'un côté, et d'autres parties prenantes telles que les autorités de réglementation et les responsables politiques de l'autre côté. La recherche indépendante d'intérêt public contribue fondamentalement à une meilleure compréhension du fonctionnement des systèmes d'IA et de leurs implications pour l'espace informationnel, y compris en matière d'éventuels risques dévastateurs. Des recherches sont également nécessaires dans des domaines moins viables économiquement, par exemple sur les systèmes de recommandation qui apportent des résultats bénéfiques à la société (voir chapitre 1, section 1.4).

Jusqu'en 2014, les modèles d'apprentissage automatique les plus importants étaient publiés par les universités. Depuis, l'industrie a pris le relais. Aujourd'hui, les investissements dans l'IA et le développement de modèles d'IA sont essentiellement réalisés par le secteur privé.[272] Ainsi, en 2020, 84 % des investissements dans l'IA provenant du secteur privé dans l'UE, contre 16 % seulement pour le secteur public, dont 30 % pour la recherche et le développement.[273] Il est donc indispensable d'accroître les investissements publics dans la recherche et le développement d'intérêt général.[274]

RECOMMANDATIONS AUX **ÉTATS**

> **Mettre en place et financer un organisme de recherche indépendant sur l'IA, national ou supranational, composé de plusieurs laboratoires de recherche indépendants. Cet organisme de recherche sur l'IA serait chargé de :**

♦ Contrôle le développement de l'IA, étudier les risques potentiels et former les législateurs et les responsables politiques.

♦ Procéder à des analyses causales plus approfondies que celles des audits traditionnels, afin d'obtenir des informations essentielles sur les effets des systèmes d'IA actuels.

271 La proposition de règlement de l'UE sur l'IA prévoit l'établissement d'un forum consultatif chargé de conseiller le Conseil et la Commission européenne sur les questions liées à l'IA (Article 58a). De plus, un panel scientifique composé d'experts indépendants doit être créé (Article 58b).

272 Stanford University Human-Centered Artificial Intelligence (2023). Artificial Intelligence Index Report 2023. Disponible sur : https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Consulté le 7 février 2024).

273 Commission Européenne (2022). AI Watch: Estimating AI Investments in the European Union. Disponible sur : https://ai-watch.ec.europa.eu/publications/ai-watch-estimating-ai-investments-european-union_en (Consulté le 7 février 2024).

274 Reconnaissant l'importance de la recherche de l'intérêt public et de la recherche pour soutenir efficacement la mise en œuvre de la législation, l'UE a établi le Centre européen pour la Transparence Algorithmique afin de soutenir l'application du règlement de l'UE sur les services numériques.
Commission européenne (2022). Centre européen pour la Transparence Algorithmique. Disponible sur : https://algorithmic-transparency.ec.europa.eu/about_en (Consulté le : 7 février 2024).

- ♦ Évaluer l'efficacité des réglementations existantes conçues pour faire face aux risques que présente l'IA, et pour développer, déployer et utiliser l'IA de façon éthique. Ces recherches seraient à la base de recommandations politiques et réglementaires ultérieures.
- ♦ Développer des alternatives aux systèmes d'IA à but lucratif qui soient au service de l'intérêt public.

> Mettre en place, financer et pérenniser un groupe d'experts issus de la société civile et de chercheurs indépendants afin de faciliter la mise en œuvre des réglementations en matière d'IA sous tous leurs aspects, y compris le contrôle, le conseil et l'évaluation des risques.[275]

> Mettre en place des systèmes de financement compétitifs et transparents pour favoriser l'émergence de capacités de recherche dans les pays de « la majorité mondiale », y compris des fonds internationaux, une aide publique au développement (APD) spécifique, des bourses d'études et des subventions de recherche.

> Collaborer avec l'Observatoire sur l'information et la démocratie[276] **pour mener des méta-études sur l'IA et parvenir à un consensus général sur les sujets controversés.**

1.5 JURIDICTIONS NATIONALES ET INTERNATIONALES

La rapidité avec laquelle les systèmes d'IA progressent pose également des défis majeurs aux instances nationales et internationales chargées de statuer sur les cas liés aux systèmes d'IA et, par conséquent, d'interpréter l'applicabilité de la législation existante. Les professionnels du droit, notamment les juges, les procureurs et les avocats, doivent disposer de l'expertise nécessaire pour représenter, examiner et faire valoir efficacement les demandes d'indemnisation des plaignants. Ils doivent également être en mesure de répondre aux appels interjetés par les développeurs et déployeurs d'IA concernant les décisions prises à leur encontre, et de traiter les infractions pénales commises par les développeurs et déployeurs de systèmes d'IA.

RECOMMANDATIONS
AUX **ÉTATS**

> Assurer une formation continue et approfondie des professionnels du droit, notamment des juges, des procureurs et des avocats en exercice, sur les capacités techniques et les aspects juridiques des systèmes d'IA, ainsi que sur leurs implications pour l'espace de l'information et de la communication. Cette formation devrait être conçue pour améliorer les connaissances juridiques dans ce domaine et pour assurer une connaissance de l'IA suffisante sur le plan technologique pour comprendre et anticiper d'éventuels risques et opportunités lors de l'examen des cas.

275 Bengio, Y. (2023). AI and Catastrophic Risk. Journal of Democracy. Disponible sur : https://www.journalofdemocracy.org/ai-and-catastrophic-risk/ (Consulté le 7 février 2024).

276 L'Observatoire sur l'information et la démocratie est un projet de méta-recherche qui agrège et synthétise la recherche pour offrir une évaluation périodique et globale de l'espace de l'information et de la communication. Pour plus d'informations : Forum sur l'Information et la Démocratie. Observatoire International de l'Information et de la Démocratie. Disponible sur : https://informationdemocracy.org/fr/mission-2/ (Consulté le : 7 février 2024).

> **Veiller à ce que les tribunaux nationaux et internationaux disposent de ressources financières et administratives suffisantes pour engager des experts techniques indépendants spécialisés dans l'IA, le cas échéant.**

> **Habiliter les tribunaux et procureurs mandatés de leurs pouvoirs d'enquête respectifs à exiger des développeurs et des déployeurs d'IA qu'ils divulguent des informations sur leurs systèmes d'IA. Cela inclut la demande de documents et de détails spécifiques relatifs au dossier à la demande du plaignant.**[277] En outre :

- Si la partie défenderesse (société ou entité d'IA) se montre réticente à coopérer avec le tribunal et les autorités chargées de l'application de la loi ou refuse de partager des informations, les tribunaux devraient avoir le pouvoir de lui imposer des sanctions. En outre, toute information non partagée par la partie défenderesse devrait être interprétée comme si elle avait été partagée, et favorable à la partie plaignante.
- Le pouvoir du tribunal d'exiger la divulgation devrait porter non seulement sur le procès, mais également sur les procédures préliminaires, ce qui permettrait aux citoyens de décider s'il y a lieu de poursuivre l'affaire au cours de la phase du procès.

> **Veiller à ce que les tribunaux internationaux, y compris la Cour européenne des droits de l'homme, la Cour interaméricaine des droits de l'homme, la Cour africaine des droits de l'homme et des peuples et la Cour pénale internationale, disposent de ressources suffisantes pour assurer la formation de leurs juristes et de leur personnel judiciaire afin d'améliorer leur connaissance des systèmes d'IA.** Cela implique de préparer du matériel pédagogique et d'encourager le partage d'informations entre les tribunaux afin de garantir une compréhension harmonisée des questions juridiques liées à l'IA dans tous les États participants.

277 La proposition de règlement de l'UE sur l'IA prévoit la création d'un forum consultatif pour conseiller le conseil d'administration et la Commission européenne sur les questions liées à l'IA (article 58 bis). En outre, un groupe scientifique d'experts indépendants devrait être créé (article 58 ter).

2. MISE EN PLACE DE PROCESSUS RIGOUREUX

L'évaluation de la conformité, l'octroi de licences et les audits sont des mécanismes essentiels pour garantir efficacement la conformité des systèmes d'IA avec les normes éthiques, juridiques et réglementaires.[278] Ensemble, ils constituent un cadre holistique de contrôle humain continu des systèmes d'IA.[279]

Les évaluations de conformité préalables à la mise sur le marché, qu'elles soient réalisées par l'entreprise elle-même ou par un tiers, constituent une forme de contrôle ex ante, garantissant que les systèmes d'IA sont conformes aux réglementations formelles existantes. Les processus d'octroi de licences, autre forme de contrôle ex ante, peuvent, le cas échéant, compléter la vérification par une autorité publique. Ils permettent de s'assurer que seuls des systèmes d'IA conformes à la législation sont utilisés. Quant aux audits, une forme de contrôle a posteriori, ils garantissent la conformité des systèmes d'IA sur la durée.

Il est impératif d'intégrer des mécanismes de contrôle dans le cadre de gouvernance de l'IA, surtout si nous tenons compte des vastes connaissances acquises dans d'autres secteurs en matière d'évaluation de la conformité, d'audit et d'octroi de licences (p.ex. dans les domaines de la finance, de la protection des données et des soins de santé). Cette expertise peut être exploitée efficacement pour mettre en place des processus de responsabilité rationalisés et cohérents pour les systèmes d'IA.

En vue de déterminer quels systèmes d'IA déployés dans l'espace de l'information et de la communication doivent faire l'objet d'une évaluation de la conformité, d'un octroi de licence et/ou d'un audit, il est nécessaire de les classer en fonction des risques qu'ils présentent et de leur incidence systémique sur la diffusion de contenus illicites, l'exercice des droits fondamentaux et les processus démocratiques (voir chapitre 1, section 2.2). Si la classification présentée au chapitre 2 peut fournir des indications sur les types de systèmes d'IA, leur catégorisation doit être effectuée au cas par cas par l'autorité compétente en matière d'IA, en tenant compte des facteurs décrits au chapitre 1, section 2.2, et résumés dans le tableau ci-dessous.

278 Laux, J. (2023). Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act. Disponible sur : https://doi.org/10.2139/ssrn.4377481 (Consulté le 7 février 2024).

279 La proposition de règlement de l'UE sur l'IA oblige les fournisseurs d'IA à coopérer avec les autorités nationales et à leur fournir la documentation pertinente si nécessaire (article 23). Le règlement de l'UE sur les services numériques prévoit également une obligation similaire (article 10).

Tableau 4.1. Facteurs permettant de déterminer le profil de risque des systèmes d'IA ayant une incidence sur l'espace de l'information

Catégorie de système d'IA	Risques liés à la diffusion de contenus illicites	Risques pour l'exercice des droits fondamentaux	Risques pour les processus démocratiques	Facteurs au cas par cas
Systèmes d'IA utilisés pour la vérification et la modération des contenus	**Oui**	**Oui**	**Oui**	• Objectif recherché • Capacité à provoquer des préjudices • Capacité à réagir aux effets négatifs et à les corriger • Transparence • Sécurité et fiabilité • Accessibilité au grand public • Accessibilité aux agents malveillants • Nombre d'utilisateurs réels et potentiels • Utilisation par les parties prenantes majeures (p.ex. médias, gouvernement) • Type et quantité de données de formation • Capacité à agir de manière autonome • Historique des préjudices causés
Systèmes d'IA utilisés pour la curation et la recommandation de contenu	**Oui**	**Oui**	**Oui**	
Systèmes d'IA utilisés pour le ciblage et la diffusion des annonces publicitaires	**Oui** (micro-ciblage)	**Oui** (micro-ciblage)	**Oui** (micro-ciblage)	
Systèmes d'IA utilisés pour la création de contenu	**Oui**	**Oui**	**Oui**	
Systèmes d'IA utilisés pour la personnalisation du contenu	**Oui**	**Oui**	**Oui**	

En principe, la classification suivante devrait s'appliquer (voir le tableau 4.2) :

- Malgré la faible probabilité de préjudice qu'ils représentent, les systèmes d'IA à faible risque devraient faire l'objet d'une évaluation de conformité interne afin de s'assurer qu'ils respectent les normes, lois, règles, réglementations et principes établis démocratiquement, notamment en matière de protection des données, de respect de la vie privée, d'atténuation des risques de biais et de non-discrimination.
- Les systèmes d'IA à risque moyen devraient non seulement faire l'objet d'évaluations de conformité ex ante en interne et par une tierce partie, menée par des organisations indépendantes accréditées par l'État, mais aussi d'un audit a posteriori tous les un à cinq ans, en fonction des exigences fixées par l'autorité chargée de l'IA.
- Les systèmes d'IA à haut risque devraient faire l'objet d'évaluations de conformité ex ante en interne et par une tierce partie, menée par des organisations indépendantes accréditées par l'État, ainsi que d'un audit annuel. Les États, en fonction de leur cadre juridique et des préjudices potentiels, pourraient également envisager la mise en place de procédures d'octroi de licence pour les systèmes d'IA à haut risque.
- Les systèmes d'IA conçus uniquement à des fins de recherche d'intérêt public, développés par des institutions de recherche d'intérêt public légitimes et approuvés en tant que projet d'intérêt public, doivent faire l'objet d'une évaluation de la conformité en interne.
- Enfin, certains systèmes et pratiques d'IA peuvent être considérés comme interdits, tels que le micro-ciblage axé sur des caractéristiques protégées.[280]

280 La proposition de règlement de l'UE sur l'IA oblige les fournisseurs de systèmes d'IA à haut risque à mettre en place des systèmes de gestion de la qualité dans le but d'assurer la conformité juridique globale des systèmes d'IA avec les réglementations juridiques, tout au long du cycle de vie des systèmes d'IA (article 17).

Cette approche exhaustive garantit que tous les systèmes d'IA, quel que soit leur niveau de risque, font l'objet d'une évaluation et d'un suivi rigoureux, assurant ainsi la conformité de leur fonctionnement aux valeurs démocratiques et à la protection des droits humains.

Tableau 4.2. Profil de risque du système d'IA et exigences en matière de contrôle*

	Évaluation de la conformité		Audit
	Interne	Externe	
Risque faible	x		
Risque moyen	x	x	x
Risque élevé*	x	x	x

* Pour les systèmes à haut risque, certains pays peuvent décider de remplacer l'évaluation de la conformité par un tiers par une procédure d'octroi de licence.

RECOMMANDATIONS AUX **ÉTATS**

> **Confier à l'autorité chargée de l'IA le soin d'établir, de publier et d'appliquer uniformément des règles relatives à la classification des systèmes d'IA dans des catégories de risque faible, moyen et élevé, ainsi qu'aux pratiques interdites. Ces dispositions devraient faire l'objet de révisions périodiques afin de s'adapter aux avancées technologiques et à leurs applications dans l'espace de l'information et de la communication.**

2.1 ÉVALUATIONS DE CONFORMITÉ

L'évaluation de la conformité avant la mise sur le marché est une forme ex ante de contrôle humain. Son principal objectif est de vérifier que les systèmes d'IA sont conformes aux normes techniques, éthiques et juridiques en vigueur. Elle peut être réalisée par les déployeurs d'IA eux-mêmes (évaluation de la conformité interne) ou par des organismes indépendants accrédités par l'État (évaluation de la conformité externe).

Actuellement, plusieurs systèmes d'IA à haut risque, tels que les systèmes intégrés dans les dispositifs médicaux, doivent faire l'objet d'évaluations de conformité en vertu des lois existantes sur la sécurité des produits. Ces évaluations sont une condition préalable à leur déploiement, ou sont requises lorsque des modifications importantes pourraient affecter la conformité du système.

De même, les systèmes d'IA conçus pour l'espace informationnel, tels que ceux utilisés par les plateformes de réseaux sociaux ou pour générer de l'information ou du contenu, devraient également faire l'objet d'évaluations de conformité rigoureuses. Bien que tous les systèmes d'IA doivent faire l'objet d'évaluations de conformité internes, les systèmes d'IA à risque moyen et les systèmes d'IA à risque élevé doivent aussi être soumis à des évaluations externes réalisées par des organismes agréés. Dans les deux cas, les résultats de ces évaluations doivent être publiés dans un rapport public (voir section 4.1).

Cette approche a pour objectif de trouver un juste équilibre entre les risques liés aux systèmes d'IA et la pression économique pouvant peser sur les développeurs et les utilisateurs de systèmes d'IA. Dans les juridictions où les systèmes d'IA à haut risque sont soumis à une procédure d'autorisation, les déployeurs ne devraient pas être contraints de se soumettre également à une procédure d'évaluation de la conformité par un tiers, afin d'éviter les surcoûts et le double emploi.

Évaluer la conformité en interne évite des coûts supplémentaires aux déployeurs, mais les oblige à contrôler de manière indépendante le respect des réglementations en vigueur. Faire évaluer les systèmes d'IA à risque moyen et élevé par des parties tierces garantit une vérification indépendante de la conformité d'un système d'IA aux normes, ce qui renforce la confiance des utilisateurs et des partenaires.

En règle générale, les évaluations de conformité doivent déterminer si les systèmes d'IA sont conformes aux réglementations nationales en matière de propriété intellectuelle, de protection des données et de droits à la vie privée, de droit de la responsabilité civile, de règles spécifiques à l'IA et de normes de cybersécurité, etc. En outre, pour garantir l'équité et la responsabilité des systèmes d'IA, les évaluations de conformité des systèmes d'IA devraient prendre en considération les principes inscrits dans les recommandations internationales telles que la Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle[281] et les principes de l'OCDE relatifs à l'IA.[282]

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer des évaluations de conformité internes pour tous les systèmes d'IA destinés à être utilisés dans l'espace de l'information et de la communication.**

> **Imposer des évaluations de conformité externes pour les systèmes d'IA à risque moyen et élevé destinés à être utilisés dans l'espace de l'information et de la communication.[283]**

> **Veiller à ce que les développeurs et déployeurs d'IA donnent accès à leurs systèmes aux organismes tiers d'évaluation de la conformité afin qu'ils puissent l'effectuer efficacement.**

> **Veiller à ce que le personnel chargé d'effectuer les évaluations de conformité, tant en interne qu'en externe, dispose des compétences nécessaires en matière de droits humains, de normes et de principes internationaux, ainsi que sur les questions relatives aux systèmes d'IA destinés à être utilisés dans l'espace informationnel.**

> **Mettre en place des dispositifs internationaux pour améliorer l'acceptation globale des résultats de tests produits par des organismes d'évaluation de la conformité (OEC) tiers certifiés, où qu'ils soient situés, ce qui inclut les accords de reconnaissance mutuelle.[284]**

281 UNESCO (2021). Recommandation sur l'éthique de l'intelligence artificielle. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Consulté le 7 février 2024).
282 OCDE (2019). Recommandation du Conseil sur l'intelligence artificielle. Disponible sur : https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449 (Consulté le 7 février 2024).
283 La proposition de règlement de l'UE sur l'IA établit des évaluations de conformité obligatoires pour les systèmes d'IA à haut risque menées par des organismes notifiés désignés par un État membre (article 30-32).
284 Voir les articles 38 à 39 de la proposition de règlement de l'UE sur l'IA concernant la coordination et la reconnaissance d'organismes notifiés appartenant à des pays tiers.

> **Imposer la publication des résultats des évaluations de conformité internes et externes et leur partage avec l'autorité compétente en matière d'IA.** Ces résultats devront être publiés dans le registre public (voir section 4.1).[285]

> Demander à **l'autorité en charge de l'IA de créer un registre public des systèmes d'IA à risque moyen et élevé qui ont été évalués. Ces informations devraient également alimenter un registre supranational géré par les Nations unies.** Celui-ci renforcerait la transparence et la responsabilité au niveau mondial en permettant aux acteurs de l'IA, en particulier dans les pays dotés d'institutions plus fragiles, de surveiller et de comparer les pratiques et les performances des développeurs et des déployeurs d'IA dans les différents pays.

2.2 OCTROI DE LICENCES

L'octroi d'une licence par une autorité compétente peut permettre d'atténuer les risques et les préjudices potentiels liés aux systèmes d'IA de manière proactive, en solutionnant les problèmes identifiés avant que ces systèmes soient mis à disposition du public. Une licence peut également contraindre les développeurs et les déployeurs de systèmes d'IA à respecter systématiquement des normes de haute qualité.[286]

Néanmoins, l'octroi de licences entraîne des coûts importants, qui risquent de renforcer davantage les positions privilégiées des grandes entreprises, plus à même de disposer des ressources nécessaires pour satisfaire aux exigences complexes en matière de licences. L'octroi de licences peut également créer une charge administrative excessive et favoriser les pratiques de corruption lorsque l'État de droit n'est pas pleinement établi.

Avant de remplacer les évaluations de la conformité externe par une procédure d'octroi de licences, les États doivent soigneusement évaluer les avantages potentiels par rapport aux inconvénients. Cela suppose d'évaluer si d'autres mesures sont plus appropriées pour atteindre les résultats souhaités, en tenant compte des spécificités nationales et de la capacité administrative de l'État.

Afin d'éviter les lourdeurs administratives aux petites entreprises et de préserver la concurrence, les États qui optent pour une procédure d'octroi de licence devraient en limiter l'application aux systèmes d'IA à haut risque. Tout comme pour les évaluations de conformité externes, les systèmes d'IA conçus exclusivement pour la recherche d'intérêt public, développés par des institutions de recherche d'intérêt public légitimes et approuvés en tant que projet d'intérêt public, devraient être exemptés de l'obligation de licence afin d'encourager l'innovation dans l'industrie et la recherche en matière d'IA.

RECOMMANDATIONS AUX **ÉTATS**

> Déterminer si la mise en place d'un **système d'octroi de licence est la procédure la plus efficace et la plus appropriée pour évaluer les systèmes d'IA à haut risque avant leur déploiement. Si un tel système est adopté :**

285 La proposition de règlement de l'UE sur l'IA exige que les systèmes d'IA à haut risque soient enregistrés dans la base de données de l'UE avant d'être mis sur le marché (article 51, article 60).

286 Malgieri, G. and Pasquale, F. (2024). Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. Computer Law & Security Review, 52. Disponible sur : https://doi.org/10.1016/j.clsr.2023.105899 (Consulté le 7 février 2024).

- **Les systèmes d'IA à haut risque devraient être légalement tenus d'adhérer à un « modèle de justification et d'explication ».**[287] En d'autres termes, pour obtenir une licence, les déployeurs d'IA devraient expliquer en détail les subtilités de leur système d'IA, justifier la raison de sa mise en service et prouver qu'il est conforme aux lois et aux directives éthiques en vigueur, dans le but de démontrer que ce système respecte, entre autres, la vie privée, la protection des données, la non-discrimination, l'exactitude, la responsabilité et les exigences en matière de sécurité.
- **Mandater l'autorité chargée de l'IA du pilotage de la procédure d'octroi de licence, c'est-à-dire de définir des orientations, des recommandations et des bonnes pratiques qui serviront de cadre aux développeurs et aux utilisateurs de systèmes d'IA à haut risque au cours de la procédure d'autorisation.**[288] Au cours de la procédure en question, l'autorité chargée de l'IA devra :
  - S'enquérir de la fiabilité technique et juridique du système d'IA et solliciter des informations supplémentaires si nécessaire.
  - Évaluer de manière indépendante et directe la fiabilité, la sécurité et la conformité des systèmes d'IA à haut risque avec les normes, lois, règles, réglementations et principes démocratiquement établis, en testant leurs fonctionnalités, le cas échéant.
  - Accorder une attention particulière à la légalité des jeux de données d'apprentissage, aux contrôles internes et à l'équilibre des déploiements de systèmes d'IA, ainsi qu'aux risques potentiels posés par le système pour les droits humains et l'espace informationnel.
  - Évaluer la conformité et l'intégrité du système d'IA non seulement au regard des exigences légales, mais également de ses implications éthiques et sociétales.
- En cas d'octroi d'une licence, **la réalisation d'audits réguliers par des tiers devrait être une condition de sa conservation**. En cas de modification substantielle du système, les déployeurs doivent soumettre les changements à l'autorité chargée de l'IA pour approbation et vérification.
- L'autorité chargée de l'IA devrait constituer un registre public des systèmes d'IA à haut risque ayant fait l'objet d'une licence.
  - Les développeurs et les déployeurs de systèmes d'IA à haut risque devraient être tenus de signaler les incidents liés à l'IA[289] dans ce registre, de partager les résultats de leurs recherches internes et de présenter des rapports d'évaluation des risques ainsi que les conclusions des audits obligatoires et volontaires.
  - L'autorité chargée de l'IA devrait examiner régulièrement ces informations, évaluer si les systèmes sous licence respectent les conditions de leur licence et élaborer des formules statistiques et des paramètres permettant de mesurer la conformité, la sécurité et l'équité des systèmes d'IA.
  - L'autorité chargée de l'IA devrait définir des seuils spécifiques, dont le franchissement entraînerait une mise en garde, une suspension temporaire du système ou le retrait pur et simple de la licence.
  - Les décisions relatives à l'octroi ou au retrait d'une licence doivent pouvoir faire l'objet d'un recours juridique le cas échéant.

287 *Ibid*
288 OCDE (2022). Responsible AI licenses: a practical tool for implementing the OECD Principles for Trustworthy AI. Disponible sur : https://oecd.ai/en/wonk/rails-licenses-trustworthy-ai (Consulté le 7 février 2024).
289 La proposition de règlement de l'UE sur l'IA oblige les fournisseurs de systèmes d'IA à haut risque à signaler les incidents aux autorités de surveillance du marché de leurs pays respectifs (article 62).

> **Exempter les systèmes d'IA exclusivement conçus à des fins de recherche d'intérêt public développés par des institutions de recherche d'intérêt public légitimes et approuvés en tant que projets d'intérêt public par une autorité responsable[290] de toute obligation en matière d'octroi de licence .** Grâce à une évaluation obligatoire de la conformité en interne, veiller à ce que ces systèmes restent conformes aux objectifs de recherche prévus et empêcher toute exploitation à des fins commerciales. Cette approche doit être contingente à l'établissement d'une définition stricte de la notion d' « intérêt public » par l'autorité compétente, qui soit raisonnablement à l'abri de toute manipulation.

2.3 AUDITS[291]

Sans réglementation ni contrôles appropriés, les systèmes d'IA peuvent perpétuer et accentuer les biais, générer (et éventuellement amplifier) des éléments de mésinformation et de désinformation préjudiciables et porter atteinte aux droits humains. Les risques qu'ils présentent peuvent être systémiques et avoir des répercussions profondes sur la fiabilité de l'espace informationnel et sur la démocratie. Si les mécanismes de contrôle ex ante (tels que l'évaluation de la conformité) peuvent garantir la conformité juridique et éthique initiale des systèmes d'IA, les audits effectués par des tiers, en tant que mécanismes de contrôle a posteriori, doivent être intégrés dans le cadre holistique de la gouvernance de l'IA pour garantir que ces systèmes restent conformes aux normes éthiques et juridiques en évolution tout au long de leur cycle de vie.

Des audits réguliers garantissent un contrôle continu et le respect des normes de sécurité, des pratiques éthiques et des obligations légales. Les audits par des tiers sont une pratique courante dans les secteurs où la sécurité et la confiance du public sont primordiales, tels que l'informatique dans le cloud et la cybersécurité.[292] Les pratiques en vigueur dans ces secteurs peuvent servir de référence pour l'élaboration de normes solides et efficaces en matière d'audit de l'IA.

La pratique des audits des systèmes d'IA ne cesse d'évoluer, et aucun consensus n'a encore émergé sur ce qu'ils devraient impliquer. Comme indiqué au point 4.4, l'accès des chercheurs aux systèmes d'IA est essentiel pour leur permettre de mieux les comprendre et d'établir les principes directeurs de leur contrôle. Ainsi, pour s'assurer que les auditeurs, les développeurs/déployeurs de systèmes d'IA et les autorités publiques comprennent tous ce processus d'audit de façon similaire, son champ d'application devrait être clairement défini. En outre, la fréquence des audits obligatoires devrait être adaptée à la gravité, à la probabilité et à la réversibilité des préjudices causés par les systèmes d'IA. Tandis que les systèmes d'IA à haut risque devraient faire l'objet d'un audit annuel par un tiers, les systèmes d'IA à risque moyen devraient faire l'objet d'un audit moins fréquent, tous les ans ou tous les cinq ans par exemple, en fonction des exigences fixées par l'autorité chargée de l'IA.

290 L'autorité responsable pourrait être l'autorité responsable de l'IA dont il est question à la section 1.2 ou l'organisme de l'agrément des chercheurs décrit à la section 4.2.

291 As a mechanism of ex-post oversight, the draft EU AI Act mandates providers of high-risk AI systems to establish a post-market monitoring system, which is based on the post-market monitoring plan (Article 61). The DSA mandates very large online platforms and very large online search engines to conduct third-party audits annually (Article 37).

292 ISACA (2018). Auditing Artificial Intelligence.
Disponible sur : https://ec.europa.eu/futurium/en/system/files/ged/auditing-artificial-intelligence.pdf (Consulté le 7 février 2024).

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer un audit externe obligatoire pour les systèmes d'IA à risque moyen et élevé déployés dans l'espace de l'information et de la communication. Tandis que les systèmes d'IA à haut risque doivent faire l'objet d'un audit annuel, les systèmes d'IA à risque moyen doivent être contrôlés tous les ans ou tous les cinq ans,** suivant les exigences fixées par l'autorité chargée de l'IA.

> **Définir le champ d'application de l'audit des systèmes d'IA à risque élevé et à risque moyen. Cela implique un audit axé sur les processus et sur les incidences, ainsi que des évaluations juridiques et techniques.** Dans ce cadre, il convient d'élaborer une liste de contrôle détaillée reprenant toutes les exigences obligatoires auxquelles les systèmes d'IA doivent se conformer.[293] Cette check-list doit comprendre des évaluations portant sur les points suivants :

- La question de la légalité de la collecte, du stockage et du traitement des données à caractère personnel au regard des lois relatives à la protection de la vie privée et des données.
- Le respect des lois sur les droits d'auteur et la propriété intellectuelle.
- L'équité et la facilité d'explication des algorithmes. Cela implique d'enquêter sur la causalité, d'identifier toute opacité ou partialité dans la logique de prise de décision des algorithmes et d'identifier les préjudices potentiels qu'ils pourraient infliger à des groupes spécifiques de la population.
- La pertinence des exigences en matière de transparence.
- La fiabilité et le bon fonctionnement des mesures de cybersécurité.[294]

> **Doter les institutions d'audit tierces de l'expertise et des pouvoirs d'investigation nécessaires,** y compris :

- La capacité d'exiger des développeurs/déployeurs de systèmes d'IA la divulgation de toutes les informations pertinentes nécessaires à l'évaluation.
- La capacité de procéder à des simulations et de contrôler les systèmes d'IA en situation réelle, y compris l'accès direct aux interfaces en ligne.[295]
- La capacité de mener des entretiens de suivi avec le personnel de l'entreprise concerné.

> **Afin de garantir leur impartialité, les auditeurs tiers doivent :**

- Être habilités par une autorité officielle.
- Déclarer leurs sources de financement et leurs éventuels conflits d'intérêts avant de procéder à l'audit.

> **Définir des règles et des orientations claires sur la structuration du processus, les obligations des développeurs/déployeurs d'IA et les droits des auditeurs, ainsi que les conséquences possibles d'un manquement aux exigences en matière d'audit. Les éléments à prendre en compte sont les suivants :**

293 *Ibid*
294 Information Commissioner's Office. A Guide to ICO Audit Artificial Intelligence (AI) Audits Contents. Disponible sur : https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf (Consulté le 7 février 2024).
295 Metaxa, D. et al (2021). Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. Disponible sur : https://www.nowpublishers.com/article/Details/HCI-083 (Consulté le 7 février 2024).

- Imposer aux développeurs/déployeurs de systèmes d'IA à risque moyen ou élevé de produire une documentation détaillée et de la soumettre aux auditeurs.
- Lors de la phase de post-audit, les auditeurs devraient fournir un retour d'information détaillé aux développeurs/déployeurs de systèmes d'IA, en précisant les domaines qui doivent être traités et en fixant des délais raisonnables.
- En cas de non-respect systématique de ces exigences, les développeurs/déployeurs d'IA devraient faire l'objet de mises en garde, d'amendes et d'une suspension temporaire/définitive du système d'IA en fonction de la gravité de la défaillance.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Garantir l'auditabilité des systèmes d'IA en fournissant une documentation détaillée, claire et complète. Cette documentation doit comprendre :**

- Des informations relatives à la provenance et à la curation des jeux de données d'apprentissage.
- Des lignes directrices relatives à l'étiquetage humain pour l'apprentissage de l'IA et le red-teaming.
- L'architecture et les capacités des modèles.
- Les tests de fiabilité.[296]
- Les évaluations des risques et leurs résultats.
- Les évaluations de conformité et leurs résultats.
- Les mesures d'atténuation mises en œuvre.
- Les dispositifs de réponse interne, tels que les mécanismes de correction, de signalement et de plainte.

2.4 MÉCANISMES DE RECOURS

Il est indispensable de mettre en place des mécanismes de recours privés et publics accessibles, justes et performants pour faire appliquer la législation de manière efficace. Les mécanismes de recours privés doivent garantir que les entreprises ne respectant pas les exigences légales et causant un préjudice à des personnes ou à des groupes protégés soient tenues de réagir rapidement et de verser des compensations proportionnelles, justes et équitables aux victimes (voir chapitre 2, section 4.3). Si les plaignants ne sont pas en mesure de trouver des solutions satisfaisantes par l'intermédiaire des mécanismes internes de traitement des plaintes de l'entreprise ou de l'entité d'IA, l'affaire doit être portée devant les tribunaux nationaux. Si un plaignant a épuisé tous les recours juridiques disponibles au niveau national, il a la possibilité de saisir les tribunaux régionaux ou internationaux compétents. En alternative d'un contrôle judiciaire, un médiateur pourrait intervenir.

296 Défini ainsi : «Les tests de fiabilité sont un type de test effectué pour évaluer la capacité d'un système ou d'un composant à fonctionner correctement lorsqu'il est soumis à des entrées invalides ou inattendues, ou lorsqu'il fonctionne en dehors de ses conditions de fonctionnement spécifiées» dans GeeksforGeeks, Robustness Testing. Disponible sur : https://www.geeksforgeeks.org/robustness-testing (Consulté le : 9 février 2024).

RECOMMANDATIONS AUX **ÉTATS**

> **Mettre en place des voies juridiques claires et accessibles pour obtenir réparation.**[297] Celles-ci devraient :
  - ♦ Définir les critères sur la base desquels les tribunaux peuvent décider de ce qui constitue un préjudice (dans les écosystèmes d'information et de communication).
  - ♦ Définir ce qui constitue une preuve et quel type de preuve est nécessaire.
  - ♦ Instaurer un système comportant plusieurs voies de recours, notamment en imposant aux entreprises d'IA de mettre en place des mécanismes internes de traitement des plaintes, permettant la saisine d'une procédure judiciaire et/ou d'un médiateur.
  - ♦ Identifier les dispositifs de compensation existants.

> **Fournir des conseils clairs et facilement compréhensibles ainsi qu'une aide juridique gratuite aux citoyens concernant ces voies juridiques.**

> **Nommer un médiateur de l'IA ou renforcer une institution de médiation existante** chargée de :
  - ♦ L'examen des réclamations non tranchées afin de trouver une solution à l'amiable entre l'entreprise ou l'entité d'IA et le plaignant. Cela inclut également l'examen des rapports des lanceurs d'alerte qui ne sont pas satisfaits du mécanisme d'examen interne de l'entreprise.
  - ♦ Si aucune solution à l'amiable ne peut être trouvée, le médiateur peut porter l'affaire devant les tribunaux, où il fera office de représentant du plaignant.[298]

> **Mettre en œuvre des mesures permettant aux utilisateurs d'accéder à des mécanismes de recours collectifs,** de sorte que lorsque plusieurs utilisateurs subissent des préjudices similaires, ils puissent déposer des plaintes communes et demander une indemnisation collective.

> Si les développeurs/déployeurs d'IA ne parviennent pas à indemniser les victimes comme l'exige le tribunal, ils devraient être sanctionnés par des amendes pour non-respect des mesures correctives. Le montant de ces amendes doit être suffisamment élevé pour inciter à la conformité et tenir compte de la gravité du manquement.[299]

297 La proposition de règlement de l'UE sur l'IA accorde aux utilisateurs le droit de déposer une plainte auprès de l'autorité de surveillance du marché compétente (article 68 bis).

298 Ogunleye, I. (2022). AI's Redress Problem: Recommendations to Improve Consumer Protection from Artificial Intelligence. Disponible sur : https://cltc.berkeley.edu/publication/cltc-white-paper-ais-redress-problem/ (Consulté le 7 février 2024).

299 Selon le règlement de l'UE sur les services numériques, la Commission peut imposer aux plateformes en ligne des amendes n'excédant pas 6 % de leur chiffre d'affaires annuel mondial total de l'exercice précédent (article 74). La proposition de règlement de l'UE sur l'IA propose des amendes comprises entre 1 et 7 % du chiffre d'affaires annuel mondial (article 71). Selon le droit des ententes de l'UE, les amendes pour infraction au droit communautaire de la concurrence ne doivent pas dépasser 10 % du chiffre d'affaires annuel global de l'entreprise.

3. GARANTIR L'IMPLICATION DES PARTIES PRENANTES

Jusqu'à présent, le domaine de l'IA a connu une surreprésentation de spécialistes des technologies étant des hommes blancs cisgenre et d'intérêts corporatistes. Cette prévalence s'est faite au détriment de la diversité des identités humaines, notamment des ethnies, des nationalités, des milieux sociaux, des langues, des religions et des croyances politiques ou autres.[300] Par conséquent, les groupes sous-représentés et marginalisés, ainsi que le grand public, n'ont eu que peu d'influence sur l'orientation du développement de l'IA. Ces lacunes en matière de représentation ont conduit au déploiement de systèmes d'IA souvent mal adaptés à l'intérêt général, excessivement « nordiques » dans leur conception, exacerbant les inégalités existantes et alimentant la méfiance à l'égard des outils alimentés par l'IA.

Pour garantir le développement de systèmes d'IA fiables, équitables et socialement bénéfiques pour tous, les mécanismes de gouvernance de l'IA devraient offrir des procédures claires favorisant une participation inclusive. À cette fin, les États devraient adopter et promouvoir le principe du pluralisme épistémologique, qui implique le respect de la diversité des points de vue, des logiques, des acteurs et des problèmes. Ceci est d'autant plus pertinent que les systèmes d'IA, dans leur forme actuelle, tendent à réduire la « biodiversité épistémique » (c.-à-d. une diversité de connaissances, de points de vue et de modes de compréhension et d'interprétation du monde d'une grande richesse, dépassant les simples différences idéologiques). Si elle n'est pas prise en compte, cette tendance constitue une menace potentielle pour les fondements du pluralisme démocratique, car elle présente une perspective unique, déterminée par des algorithmes, comme indiscutablement objective.[301]

3.1. PARTICIPATION DES PARTIES PRENANTES AUX PROCESSUS DE GOUVERNANCE DE L'IA

Outre la création d'un conseil consultatif ou l'intégration directe des différents parties prenantes au sein de l'autorité chargée de l'IA, les États devraient également prévoir des mesures proactives pour permettre à la société civile, aux chercheurs, aux journalistes et au grand public de s'impliquer dans la gouvernance de l'IA.

RECOMMANDATIONS
AUX **ÉTATS**

> **Mettre en place un portail public pour recueillir les commentaires des citoyens sur les systèmes d'IA, la réglementation et la gouvernance. Ce portail serait géré par l'autorité chargée de l'IA, qui veillerait à ce que les utilisateurs ayant soumis des commentaires reçoivent un retour d'information en temps utile**.

300 Organisation des Nations Unies (2001). Déclaration universelle des droits de l'homme. Article 2. Disponible sur : https://www.un.org/fr/about-us/universal-declaration-of-human-rights (Consulté le 7 février 2024).
301 Miller, T. et al. (2008). *Epistemological Pluralism: Reorganizing Interdisciplinary Epistemological Pluralism: Reorganizing Interdisciplinary Research Research*. *Ecology and Society.* 13(2): 46. Disponible sur : https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1034&context=unf_research (Consulté le 7 février 2024).

> **Inviter et financer les OSC et les chercheurs indépendants à participer à des évaluations de conformité par des tiers** (comme indiqué à la section 2.1)**, à des processus d'audit** (comme indiqué à la section 2.3) **et à des évaluations des risques systémiques** (comme indiqué à la section 2.2 du chapitre 1).

> **Imposer aux entreprises et aux entités spécialisées dans l'IA la mise en place d'un portail de notification et de réclamation public, où chacun peut signaler les préjudices et les risques que le système d'IA pose à l'espace de l'information et de la communication.** Cette disposition est particulièrement importante en cas de violation des droits collectifs, car elle donne aux représentants de la société civile les moyens de s'engager et d'assurer la responsabilisation des systèmes d'IA et leur incidence sur la société.

> **Solliciter activement la contribution des OSC, des chercheurs, des journalistes et du grand public à l'élaboration des politiques publiques en matière d'IA, y compris les processus de réglementation, de normalisation et d'évaluation de la conformité,** conformément aux meilleures pratiques en vigueur en matière de participation des différentes parties intéressées.

> **Mettre en place des assemblées citoyennes de délibération publique sur les risques et les opportunités de l'IA pour l'espace de l'information et de la communication, ainsi que sur les options de gouvernance afin de guider les politiques et les réglementations publiques.** Ces assemblées devraient obligatoirement comprendre des représentants de minorités et d'autres groupes vulnérables ou marginalisés.

RECOMMANDATIONS
AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Mettre en place un portail de notification et de plainte public où chacun peut signaler les préjudices et les risques que pose le système d'IA pour l'espace informationnel**.

- Permettre à quiconque de consulter les plaintes concernant le système (et non les plaintes personnelles).
- Présenter un rapport annuel sur les plaintes, les avis et les retours d'information reçus, y compris les mesures prises pour y donner suite.

3.2 FINANCEMENT DES OSC

Les OSC indépendantes et d'intérêt public ont un rôle essentiel pour assurer la responsabilisation et le contrôle des entreprises. Pourtant, leur participation est souvent considérée comme allant de soi, et n'est pas rémunérée. Ces organisations ont besoin de ressources financières pour investir leur temps et leur énergie dans une participation efficace et significative, de façon régulière et pérenne. Or, cela s'avère particulièrement difficile dans les pays de « la majorité mondiale », où les OSC peuvent se heurter à des difficultés supplémentaires pour obtenir des financements. Pour permettre aux OSC de continuer à participer aux processus législatifs, réglementaires et de contrôle de l'IA, il est urgent d'établir des modèles de financement qui garantissent leur soutenabilité économique et préservent leur indépendance vis-à-vis des gouvernements et des intérêts économiques.

RECOMMANDATIONS
AUX **ÉTATS**

> **S'engager à compenser financièrement les OSC qui participent aux institutions et structures officielles de contrôle et de gouvernance de l'IA.** Des mécanismes de sélection indépendants, ainsi que de transparence et de responsabilité, doivent être mis en place pour garantir l'indépendance des OSC. Cela comprend également des mesures de protection contre les conflits d'intérêts.

> **Créer un fonds indépendant pour allouer des ressources financières aux OSC spécialisées dans l'IA, la démocratie et l'intégrité de l'information.** Garantir l'indépendance des OSC qui reçoivent ces fonds.

> **Mettre en place des mécanismes favorisant la contribution substantielle des entreprises et entités du secteur de l'IA au financement des OSC.** Il peut s'agir de consacrer un pourcentage spécifique de l'impôt sur les sociétés d'IA (voir chapitre 3, section 3.3) et un pourcentage des amendes prononcées à l'encontre des sociétés d'IA.

> **Instituer un fonds international et consacrer une partie de l'aide publique au développement (APD) au financement des OSC spécialisées dans l'IA, la démocratie et l'intégrité de l'information dans les pays de « la majorité mondiale ».**[302]

4. GARANTIR LA TRANSPARENCE ET L'ACCÈS AUX DONNÉES

L'opacité qui caractérise actuellement le secteur de l'IA est l'un des plus sérieux défis qui menacent l'écosystème démocratique de l'information et de la communication. Les systèmes d'IA ne sont généralement pas assortis d'une documentation complète, ce qui complique la compréhension de leur conception et de leur processus de développement. En outre, il existe un manque de transparence quant à la manière dont les systèmes d'IA fonctionnent et sont évalués, ce qui est pourtant essentiel à l'évaluation de leur fiabilité.

Ce manque de transparence prive les utilisateurs, les organismes de contrôle et les régulateurs d'informations essentielles sur les systèmes d'IA, y compris sur les divers facteurs qui les affectent. Il empêche également de vérifier si les développeurs et les déployeurs respectent les principes et les normes éthiques. De plus, cette opacité compromet la capacité des chercheurs à déterminer la causalité et à contribuer à l'accumulation de connaissances sur les risques et les perspectives que présentent les systèmes d'IA. Enfin, elle prive les responsables politiques des informations indispensables à l'élaboration de réglementations efficaces.

302 OCDE (2023). L'Aide Publique au Développement (APD) - OCDE. Disponible sur : https://www.oecd.org/fr/cad/financementpourledeveloppementdurable/normes-financement-developpement/aide-publique-au-developpement.htm (Consulté le 7 février 2024).

4.1 DIVULGATION DES SYSTÈMES ET DES DONNÉES D'IA

Comme indiqué au chapitre 2, section 1.2, il est primordial d'adopter une approche de la transparence à plusieurs niveaux pour remédier à l'opacité qui prévaut actuellement autour des systèmes d'IA et favoriser la responsabilisation. Il convient donc de veiller à ce que les informations essentielles soient accessibles aux différents parties prenantes d'une manière qui tienne compte de leurs besoins, objectifs, fonctions et capacités d'interprétation spécifiques.

Dans ce contexte, les États jouent un rôle déterminant en établissant des exigences minimales en matière de transparence pour les entreprises et entités spécialisées dans l'IA, leurs systèmes et les données qu'elles utilisent.

Cette section se concentre exclusivement sur les exigences de transparence pour les systèmes d'IA générative.[303] Pour une analyse détaillée des exigences de transparence applicables aux plateformes et aux systèmes d'IA traditionnels qu'elles utilisent, consultez le rapport du Forum sur l'information et la démocratie intitulé *Pour mettre fin aux infodémies* (2020).[304]

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux développeurs et déployeurs d'IA d'assurer la transparence de leurs systèmes selon une approche graduelle, en fournissant des informations au grand public et des informations plus détaillées aux autorités de régulation et aux chercheurs agréés.**
> **Les informations communiquées au grand public et conservées dans un registre public doivent comprendre :[305]**
> - ♦ Des informations générales sur les algorithmes, notamment :
>   - • Une vue d'ensemble de leurs fonctionnalités, de leurs principaux objectifs, de ce qui justifie leur processus décisionnel, et spécification des données entrantes et sortantes.
>   - • Les outils de contrôle humain.
>   - • L'évaluation des biais et de l'équité en soulignant leur incidence sur les groupes marginalisés.[306]
> - ♦ Des informations générales sur les capacités et les limites des systèmes d'IA : leurs fonctions, leur potentiel, les cas d'utilisation recommandés et les dangers possibles d'une mauvaise utilisation.
> - ♦ Des informations générales sur la contrôlabilité du système : le degré de fiabilité et d'adéquation d'un système avec ses objectifs initiaux, la probabilité et les indicateurs possibles d'un préjudice involontaire causé par le système, et les mesures pouvant être prises si le système devenait incontrôlable.

303 Bell, A. et al (2023). Algorithmic Transparency Playbook. Center for Responsible AI. Disponible sur :https://dataresponsibly.github.io/algorithmic-transparency-playbook/resources/transparency_playbook_camera_ready.pdf (Consulté le 7 février 2024).
304 Forum sur l'information et la démocratie (2020). Pour mettre fin aux Infodémies. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Infodemies_FR.pdf (Consulté le 8 février 2024).
305 Le règlement de l'UE sur les services numériques impose aux plateformes en ligne de publier régulièrement des rapports de transparence (article 42). En outre, la Commission peut, sur simple demande, exiger des plateformes en ligne qu'elles fournissent les informations relatives à l'infraction présumée (article 67), procéder à des entretiens (article 68) et effectuer des inspections dans leurs locaux (article 69).
306 OCDE. Algorithmic Transparency Recording Standard. Observatory of Public Sector Innovation. Disponible sur : https://oecd-opsi.org/innovations/algorithmic-transparency-standard/ (Consulté le 7 février 2024).

- Des informations générales sur les incidences du système, par exemple, sur les effets étudiés et potentiels du système d'IA sur la dynamique sociopolitique et les droits humains ou sur les mesures de protection supplémentaires à adopter par les consommateurs pour utiliser le produit de manière plus sûre.[307]
- Les résultats de l'évaluation des risques.
- La taille, la composition, la portée et l'ampleur des tâches humaines utilisées pour la formation des jeux de données.
- Les pratiques de curation de la provenance des données utilisées pour la formation des jeux de données (voir chapitre 1, section 1.a).
- Les caractéristiques techniques générales du modèle, les capacités de calcul et son empreinte sur l'environnement.[308]
- Les lignes directrices utilisées pour le red-teaming et l'étiquetage humain pour l'apprentissage de l'IA.

> **Les informations transmises aux régulateurs et aux chercheurs indépendants agréés doivent être facilement accessibles et exploitables à des fins de recherche, et doivent être incluses dans un registre centralisé** (au même titre que les informations fournies au grand public, comme décrit ci-dessus) :

- Accès au « modèle de base », c.-à-d. aux versions du modèle avant et après la mise au point, aux interfaces de programmation d'applications (API), aux informations sur les familles de systèmes et aux éléments internes du système (p.ex., les métadonnées).[309]
- Les méthodologies et les résultats de l'évaluation des risques.
- Des informations détaillées sur les données d'apprentissage.
- La documentation technique, y compris les mesures de performance, les algorithmes et les processus de validation des systèmes d'IA.
- Les conclusions des recherches sur l'efficacité des mesures d'atténuation des risques pour protéger les utilisateurs contre les préjudices et les violations des droits humains
- Les résultats des recherches internes ainsi que les méthodologies utilisées pour les mener à bien.

4.2 ACCÈS AUX DONNÉES À DES FINS DE RECHERCHE INDÉPENDANTE

Des recherches indépendantes sur les incidences des systèmes d'IA sont essentielles pour renforcer la transparence, la responsabilisation et la sécurité, et, en définitive, pour garantir une gouvernance démocratique.

À cette fin, il est nécessaire d'obliger les développeurs et les déployeurs de systèmes d'IA à prévoir des modalités claires permettant aux chercheurs d'accéder aux données relatives à l'apprentissage, au déploiement et à l'utilisation de ces systèmes.

307 Anderljung, M. et al (2023). Towards Publicly Accountable Frontier LLMs. Disponible sur : https://arxiv.org/pdf/2311.14711.pdf (Consulté le 7 février 2024).
308 Bommasani, R. et al (2023). The Foundation Model Transparency Index. Disponible sur : https://crfm.stanford.edu/fmti/fmti.pdf (Consulté le 7 février 2024).
309 Anderljung, M. et al (2023). Towards Publicly Accountable Frontier LLMs. Disponible sur : https://arxiv.org/pdf/2311.14711.pdf (Consulté le 7 février 2024).

Cependant, plusieurs garanties doivent être apportées pour protéger simultanément les intérêts du public, les institutions démocratiques, la vie privée des utilisateurs, l'indépendance des chercheurs et les secrets commerciaux des entreprises. En premier lieu, les États doivent mettre en place ou désigner un organisme de contrôle indépendant chargé d'examiner les demandes de recherche. Cet organisme doit ensuite mettre en place et gérer un mécanisme de hiérarchisation permettant d'évaluer et de classer stratégiquement les projets de recherche en fonction de leur valeur scientifique et de leur cohérence avec les priorités politiques et les intérêts de la société. Enfin, les États doivent imposer des garanties claires en matière de cybersécurité et de protection de la vie privée aux développeurs et aux utilisateurs de systèmes d'IA, ainsi qu'aux chercheurs indépendants agréés, garantissant ainsi le respect des droits fondamentaux des utilisateurs.

Pour des recommandations détaillées sur l'accès des chercheurs aux plateformes et leur fonctionnement, consultez le rapport du Forum sur l'information et la démocratie relatif à la transparence des plateformes et à l'accès aux données à des fins de recherche intitulé *Pour mettre fin aux infodémies* (2020).[310]

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux développeurs et déployeurs d'IA d'accorder l'accès aux données des systèmes d'IA à des chercheurs indépendants agréés, sur demande motivée d'un organisme indépendant** (voir section 4.1).

> **Instituer ou désigner un organisme national indépendant[311] compétent pour :**

- **Déterminer si un chercheur souhaitant accéder aux données des systèmes d'IA répond aux critères d'éligibilité définis et lui accorder le statut de chercheur indépendant agréé.** Pour pouvoir bénéficier du statut de chercheur agréé, un chercheur doit :
  - Être affilié à une institution de recherche reconnue, ce qui inclut les universités, les bibliothèques, les instituts de recherche, les groupes de réflexion, les organisations de la société civile ou toute autre entité indépendante dont l'objectif principal est la recherche scientifique ou le développement d'activités éducatives sans but lucratif.
  - Révéler les sources de financement de ses recherches.
  - Démontrer sa capacité à respecter les exigences en matière de sécurité et de confidentialité des données nécessaires à la protection des données à caractère personnel et à l'intégrité de l'information.
  - Démontrer que ses demandes d'accès aux données et les délais associés sont nécessaires et proportionnés.
  - S'engager à mettre gratuitement à la disposition du public les résultats de ses recherches.

310 Forum sur l'information et la démocratie (2020). Pour mettre fin aux Infodémies. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Infodemies_FR.pdf (Consulté le 8 février 2024).

311 Le règlement de l'UE sur les services numériques définit déjà un cadre qui permet aux chercheurs d'intérêt public agréés d'accéder aux données des très grandes plateformes (c'est-à-dire les plateformes en ligne comptant plus de 45 millions d'utilisateurs actifs dans l'UE) sous certaines conditions. Ce cadre pourrait être adapté pour faciliter l'accès aux données pertinentes pour la recherche sur les systèmes d'IA. Les coordinateurs nationaux des services numériques (CSN) et le coordinateur européen des services numériques (CESN) pourraient être chargés d'examiner les applications de recherche liées aux systèmes d'IA déployés dans l'UE (article 40, paragraphes 9 et 10).

- **Examiner les demandes de recherche soumises par des chercheurs indépendants agréés à des entreprises et entités du secteur de l'IA et, le cas échéant, adresser une demande motivée d'accès aux données à l'entreprise ou à l'entité en question. Lors de l'examen des demandes de recherche reçues, cet organisme devrait :**
  - **Évaluer leur valeur scientifique.** Une demande de recherche ne doit être approuvée que si elle est d'intérêt public et utilisée à des fins non commerciales. Les demandes de recherche doivent de préférence être consacrées aux capacités, à la contrôlabilité et aux incidences des systèmes d'IA.[312]
  - **Classer les demandes de recherche par ordre de priorité en fonction de leur faisabilité, de leur unicité, de leur pertinence sociopolitique et de leur intérêt public.**
- **Imposer aux développeurs/déployeurs de systèmes d'IA de garantir une cybersécurité et une protection de la vie privée appropriées, auxquelles les chercheurs indépendants agréés doivent se conformer lorsqu'ils reçoivent, manipulent ou analysent des données pertinentes.**
- **Proposer aux chercheurs indépendants agréés des conseils sur les mesures nécessaires en matière de protection de la vie privée et de cybersécurité et les aider à se conformer aux normes en matière de cybersécurité et de protection de la vie privée.**

RECOMMANDATIONS
AUX DÉVELOPPEURS ET DÉPLOYEURS DE SYSTÈMES D'IA

> **Autoriser les chercheurs indépendants agréés à accéder aux données de l'IA pour mener des recherches dans l'intérêt public** (voir section 4.1) **et prévoir des mécanismes permettant d'assurer cet accès sans compromettre la confidentialité des données, la propriété intellectuelle et la sécurité.**

> **Prévenir les utilisateurs que leurs données sont partagées à des fins de recherche en publiant des avis ou par d'autres moyens appropriés, et les informer sur les mesures de protection de leur vie privée et sur le type d'informations partagées.** Les utilisateurs dont le profil ne contient pas de contenu public doivent avoir la possibilité de refuser que leurs informations soient partagées avec un chercheur indépendant agréé.

4.3 ÉVALUATIONS EXPÉRIMENTALES SUR LES PLATEFORMES

Il est indéniable que les tests A/B (à savoir la méthode d'évaluation expérimentale visant à comparer deux versions similaires d'un produit afin de comparer leurs performances de manière contrôlée)[313] constituent la méthodologie la plus efficace pour établir la causalité lors de l'étude d'impact des systèmes

312 Anderljung, M. et al(2023). Towards Publicly Accountable Frontier LLMs. Disponible sur : https://arxiv.org/pdf/2311.14711.pdf (Consulté le 7 février 2024).
313 Gallo, A. (2017). A Refresher on A/B Testing. Harvard Business Review. Disponible sur : https://hbr.org/2017/06/a-refresher-on-ab-testing (Consulté le 7 février 2024).

d'IA sur le comportement de l'utilisateur. Plus encore que les méthodologies fondées sur l'observation des données, les études A/B sont un mécanisme très performant pour vérifier les incidences directes de ces systèmes.

Dans la mesure où les capacités de test A/B sont généralement propriétaires et limitées aux entreprises propriétaires du système, la collaboration entre les chercheurs externes et le secteur privé est fondamentale.

À l'heure actuelle, aucune juridiction n'impose aux entreprises et entités spécialisées dans l'IA d'autoriser des chercheurs externes indépendants à mener ce type d'intervention expérimentale.[314]

Si quelques exemples de collaboration volontaire existent entre des entreprises et entités spécialisées dans l'IA et des chercheurs indépendants[315], les États doivent impérativement définir des obligations légales afin de garantir que les expériences menées dans l'intérêt du public puissent être menées indépendamment du pouvoir discrétionnaire des dirigeants de l'industrie des technologies.

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux développeurs et déployeurs d'IA de permettre à des chercheurs indépendants agréés de procéder à des évaluations expérimentales des systèmes d'IA, sur demande motivée d'un organisme indépendant.**

> **Charger l'organisme (inter)national indépendant mentionné à la section 4.2 de définir des critères d'éligibilité et de hiérarchisation pour la réalisation d'évaluations expérimentales de systèmes d'IA, en coopération avec des universitaires, des représentants de la société civile et d'autres parties prenantes concernés par l'IA.**

> **Imposer aux chercheurs indépendants agréés l'obligation légale de signer un accord de non-divulgation des informations confidentielles des utilisateurs et de l'avantage concurrentiel des entreprises.**

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Autoriser les chercheurs indépendants agréés à mener des évaluations expérimentales des systèmes d'IA dans l'intérêt public et prévoir des mécanismes clairs pour mener ces évaluations d'une manière qui ne compromette pas la confidentialité des données, la propriété intellectuelle et la sécurité.**

314 L'inclusion potentielle des tests A/B dans le champ d'application des pouvoirs à accorder aux chercheurs externes dans le cadre du règlement de l'UE sur les services numériques reste incertaine. Cet aspect est actuellement en discussion et fait l'objet d'efforts de plaidoyer permanents. Le règlement délégué qui finalisera ces détails est encore en cours d'élaboration.
Cependant, il est important de noter que l'acte délégué déjà adopté et relatif aux audits menés dans le cadre du règlement de l'UE sur les services numériques accorde aux auditeurs le pouvoir de tester les systèmes de recommandation en utilisant les méthodes appropriées, telles qu'indiquées dans les clauses 25 à 30 du document.
Commission Européenne (2023). Delegated Regulation on independent audits under the Digital Services Act. Disponible sur: https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act (Consulté le 7 février 2024).

315 Le projet de Meta Newsfeed Toronto/Berkeley, une collaboration entre Meta et des chercheurs universitaires du Center for Human-Compatible AI de Berkeley et de l'Institut Schwartz Reisman et de l'Institut Vector de l'Université de Toronto, illustre la collaboration volontaire entre une entreprise d'IA et des chercheurs indépendants. Co-dirigé par Jonathan Stray et Gillian Hadfield, le projet vise à optimiser l'algorithme de recommandation du fil d'actualités de Facebook pour d'autres mesures que celle de l'engagement des utilisateurs. Comme résumé dans : Transparency Mechanisms for Social Media Recommender Algorithms: from Proposals to Actions. Report, Global Partnership on AI. Disponible sur : https://gpai.ai/projects/responsible-ai/transparency-mechanisms-for-social-media-recommender-algorithms.pdf (Consulté le 8 février 2024).

4.4 BACS À SABLE DE RESPONSABILISATION POUR LES ALGORITHMES D'IA

Si certaines juridictions (notamment l'UE) exigent désormais des plateformes qu'elles partagent la documentation relative à leurs algorithmes et qu'elles autorisent les chercheurs à accéder aux données, les entreprises ont tendance à se montrer réticentes lorsqu'il s'agit de partager des informations plus détaillées avec des chercheurs externes, telles que le code source de leurs algorithmes. Cette réticence est souvent justifiée par des motifs légitimes, notamment la nécessité de protéger les secrets commerciaux.

Pour faciliter la responsabilisation tout en tenant compte des préoccupations des entreprises, les États devraient imposer la mise en place de « bacs à sable de responsabilisation ». Contrairement aux bacs à sable conventionnels utilisés pour le développement technologique, les bacs à sable de responsabilisation permettraient aux parties prenantes externes de tester les algorithmes et les systèmes d'IA. Ils pourraient introduire des données dans un système d'IA au sein du bac à sable et ainsi obtenir des informations sur le fonctionnement de ce système.

De ce fait, ces nouveaux environnements de test permettraient aux chercheurs indépendants, aux organisations de la société civile et aux autorités réglementaires de tenir les entreprises et les entités d'IA responsables de la conception de leurs systèmes d'IA, tout en préservant la confidentialité du code et des paramètres concernés.

RECOMMANDATIONS AUX **ÉTATS**

> **Imposer aux entreprises et entités spécialisées dans l'IA de mettre en place un «bac à sable de responsabilisation» accessible aux parties prenantes externes,** notamment les chercheurs indépendants, les organisations de la société civile et les autorités réglementaires. **Ce « bac à sable de responsabilisation » devrait donner accès aux systèmes d'IA qu'ils développent et/ou déploient par l'intermédiaire d'une interface de programmation d'applications (API),** ce qui permettrait de tester et d'analyser ces systèmes de manière externe et indépendante, tout en protégeant les informations exclusives.

RECOMMANDATIONS AUX **ENTREPRISES ET ENTITÉS D'IA**

> **Mettre en place un « bac à sable de responsabilisation » donnant accès aux systèmes d'IA développés et/ou déployés via une API pour des tests et analyses externes et indépendants.**

5. PROMOUVOIR LA COOPÉRATION ET LA RÉGLEMENTATION INTERNATIONALES

Par le biais d'une coopération et d'une réglementation internationales, la gouvernance mondiale de l'IA peut jouer un rôle déterminant dans l'évolution du paysage de l'IA.

Une telle gouvernance peut contribuer à établir des normes, des principes et éventuellement des règles de portée mondiale à obligatoirement prendre en compte lors du développement, du déploiement et de l'utilisation de systèmes d'IA dans l'espace informationnel. Ces normes et réglementations devraient être élaborées dans l'intérêt de la démocratie, de l'intégrité de l'information et du respect des droits humains des citoyens du monde entier.

La coopération internationale peut également promouvoir une gouvernance efficace de l'IA au niveau national grâce à l'apprentissage par les pairs, au partage des meilleures pratiques et à l'effet de levier international. Elle peut également soutenir le renforcement des capacités et des connaissances, en particulier dans les pays aux institutions moins bien dotées en ressources.

Il est donc essentiel de tenir compte des défis et des opportunités propres aux pays de « la majorité mondiale » pour garantir des pratiques responsables et éthiques en matière d'IA. Il faut notamment redéfinir les objectifs politiques en matière d'équité afin de traiter les questions spécifiques aux pays de « la majorité mondiale », telles que la discrimination fondée sur la caste. Par ailleurs, les régimes de confidentialité des données dans les pays de « la majorité mondiale » étant moins développés que dans les pays de l'OCDE, la coopération internationale peut aider les nations à mettre en place des régimes de confidentialité rigoureux sur lesquels les réglementations en matière d'IA peuvent s'appuyer.

Enfin, la coopération internationale peut renforcer les capacités des pays de « la majorité mondiale ». Cela suppose de renforcer leurs capacités techniques et de favoriser la compréhension de l'impact des systèmes d'IA sur ces pays, ainsi que de formuler des réglementations adaptées à leur contexte spécifique.

Comme l'indique le rapport de l'organe consultatif de haut niveau sur l'IA du secrétaire général des Nations unies, la gouvernance internationale de l'IA devrait reposer sur des principes inclusifs et être guidée par l'intérêt public et le droit international des droits humains. Parmi les fonctions d'une structure de gouvernance internationale figurent la recherche, le renforcement de l'interopérabilité et l'élaboration de normes, la facilitation du déploiement de l'IA au profit de la société, la collaboration sur les jeux de données, les systèmes et les talents en matière d'IA, la surveillance des risques et l'intervention en cas d'urgence, ainsi que la définition de normes.[316]

Tandis que diverses organisations internationales, dont les Nations unies, l'UNESCO, le G7, l'OCDE, l'OSCE et le Conseil de l'Europe, ont entrepris d'établir des principes universels et de mettre en place une structure internationale de gouvernance de l'IA, ces initiatives ne sont pas toujours en phase ou complémentaires les unes avec les autres. Pour optimiser leur efficacité, une plus grande harmonisation et une collaboration plus active entre ces organisations sont nécessaires.

Enfin, les pays démocratiques devraient renforcer leur coopération en s'appuyant sur le Partenariat pour l'information et la démocratie afin d'établir des bonnes pratiques et des normes et principes

316 Organce consultatif de haut niveau du Secrétaire général de l'ONU sur l'IA (2023). Interim Report: Governing AI for Humanity. Disponible sur : https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/ai_advisory_body_interim_report.pdf (Consulté le : 7 février 2024).

mondiaux concernant les systèmes d'IA et leur impact sur l'espace informationnel. Un forum mondial sur l'IA inspiré de la structure de gouvernance de l'Internet Corporation for Assigned Names and Numbers (ICANN) pourrait servir de lieu de dialogue ouvert à tous et constituer une plaque tournante pour les questions d'IA telles que la recherche, en coopération avec l'Observatoire sur l'information et la démocratie[317], la facilitation de l'harmonisation des initiatives internationales et la promotion de principes et d'une législation internationaux. Une telle initiative devrait s'appuyer sur les efforts existants tels que le Partenariat mondial sur l'intelligence artificielle[318] ou le Réseau d'experts de l'OCDE sur l'IA[319], tout en garantissant que la société civile, les médias et les journalistes, ainsi que d'autres représentants de l'intérêt public et de la communauté, disposent de la même légitimité à la table des négociations.

RECOMMANDATIONS AUX **ÉTATS**

> **Veiller à ce que la gouvernance internationale de l'IA soit régie par des principes démocratiques en renforçant la coopération dans le cadre du Partenariat pour l'information et la démocratie afin de promouvoir la recherche, de développer les meilleures pratiques et d'établir des normes et des principes généraux pour le développement, le déploiement et l'utilisation des systèmes d'IA dans l'espace informationnel.**

> **Veiller à ce que la gouvernance de l'IA repose sur les valeurs de justice, d'équité et de non-discrimination.** À cette fin, **il est essentiel de répondre de manière adéquate aux besoins des pays de « la majorité mondiale »,** notamment par les actions suivantes :

- Redéfinir les objectifs politiques relatifs à la notion d'équité afin qu'ils englobent les questions spécifiques à « la majorité mondiale » (p.ex. la discrimination fondée sur la caste).
- Prendre des mesures visant à lutter contre les inégalités en termes d'accès aux connaissances et aux ressources pour le développement et le déploiement des systèmes d'IA.
- Soutenir la mise en place de cadres réglementaires solides, par exemple en matière de protection de la vie privée.

> **Promouvoir, en s'appuyant sur les initiatives existantes, la création d'un Forum mondial de l'IA pour un dialogue ouvert, avec une participation soutenue et égale de la société civile, des médias et des journalistes, des chercheurs et d'autres organisations communautaires et d'intérêt public. L'objectif d'un tel forum est de faciliter un échange ouvert entre les différentes parties prenantes sur les questions clés de la gouvernance de l'IA.** Au fil du temps, ce forum devrait évoluer vers un système de gouvernance plus structuré, mais une telle configuration dynamique garantirait que sa structure reste adaptable et inclusive à long terme, afin de répondre efficacement aux progrès rapides et aux défis complexes posés par les technologies de l'IA. Dans cette optique, les États devraient s'inspirer d'organisations telles que l'ICANN (Internet Corporation for Assigned Names and Numbers) pour une coordination efficace de la gestion des ressources de l'Internet, ainsi que des pratiques des organismes internationaux de normalisation.

317 Forum sur l'information et la démocratie. Observatoire International sur l'information et la démocratie. Disponible sur : https://informationdemocracy.org/fr/mission-2/ (Consulté le 7 février 2024).
318 The Global Partnership on Artificial Intelligence (2020). About - GPAI. Disponible sur : https://gpai.ai/about/ (Consulté le 7 février 2024).
319 OCDE. OECD Working Party and Network of Experts on AI. Disponible sur : https://oecd.ai/fr/network-of-experts (Consulté le 7 février 2024).

> Un tel Forum mondial de l'IA pour un dialogue ouvert pourrait :

- **Servir de source centrale pour la recherche et le développement en matière d'IA, offrir une expertise et des conseils sur les questions liées à l'IA aux entités nationales** impliquées dans la gouvernance d'IA, y compris les systèmes juridiques et judiciaires. Il coopérerait étroitement avec l'Observatoire sur l'information et la démocratie pour mener des métarecherches sur l'impact de l'IA sur l'espace informationnel.
- **Favoriser l'harmonisation des initiatives mondiales en matière d'IA, encourager la convergence et contribuer à l'élaboration et à la diffusion de normes mondiales relatives à l'IA.**
- **Afin de renforcer l'interopérabilité, développer et promouvoir des principes et des bonnes pratiques en matière de législation et d'élaboration de normes, tels que ceux relatifs à la provenance des données et aux pratiques de curation** (voir chapitre, section 1.a).
- **Collaborer avec des organismes internationaux** pour relever les défis de l'IA dans le cadre des conventions existantes, telles que la Convention de Budapest sur la cybercriminalité.
- **Suivre les tendances en matière de développement et de réglementation de l'IA, surveiller les risques et anticiper l'impact futur, en qualité d'organisme de contrôle international.**
- **Favoriser la coopération pour concevoir et partager des alternatives publiques aux systèmes, jeux de données et infrastructures d'IA à but lucratif ; développer les talents et mener des campagnes de sensibilisation à l'IA** (voir chapitre 3, section 1.5 et section 3.2).

> **Prendre la mesure et intégrer l'importance de la promotion du développement et du déploiement de systèmes d'IA éthiques sur le plan commercial. À cette fin, on pourra :**

- **Intégrer des clauses éthiques sur l'IA dans les accords commerciaux.** Ces derniers devraient inclure des clauses favorisant le développement et le déploiement d'une IA éthique, en s'inspirant de la *Recommandation sur l'éthique de l'intelligence artificielle* de l'UNESCO.[320] Les clauses relatives à l'IA éthique pourraient inclure :
  - L'interdiction des pratiques d'IA discriminatoires fondées sur des données sensibles.
  - La coopération en matière de renforcement des capacités et de partage des connaissances concernant les pratiques éthiques en matière d'IA afin de créer des conditions de concurrence équitables, en privilégiant les pays de « la majorité mondiale ». Cela pourrait se traduire par des initiatives de recherche communes et des accords de transfert de technologies.
- **Coopérer avec les organismes internationaux de normalisation afin de définir des normes en matière d'IA éthique qui pourraient être intégrées aux accords commerciaux,** fournissant ainsi un cadre cohérent entre les différentes juridictions.

320 UNESCO (2021). Recommandation sur l'éthique de l'intelligence artificielle. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Consulté le 7 février 2024).

REMERCIEMENTS

Le Forum tient à remercier chaleureusement les membres de ce Groupe de travail, en particulier ses co-présidents et les 12 autres experts qui ont accepté de le rejoindre, pour leur temps, leur point de vue éclairant et leur soutien tout au long de ce processus.

Le Forum tient à exprimer sa gratitude envers les représentants des États et des institutions publiques du Partenariat pour l'information et la démocratie qui ont participé à la réunion de consultation du 28 novembre 2023, et les remercier pour leurs idées et leurs observations sur les recommandations préliminaires.

Le Forum tient également à remercier chaleureusement les près de 40 experts que nous avons interviewés, pour leur temps, leurs perspectives éclairantes et leur soutien.

- **Ginny Badanes,** *Democracy Forward, Microsoft*
- **Yoshua Bengio,** *Professeur d'université, Université de Montréal; Fondateur et Directeur Scientifique, Mila, Institut québécois d'intelligence artificielle*
- **Renato Berrino Malaccorto,** *Open Data Charter*
- **Jamie Berryhill,** *Analyste en Intelligence Artificielle, OCDE*
- **Bruno Bioni,** *Fondateur et Directeur, Data Privacy Brasil*
- **Lena-Maria Böswald,** *Responsable de politique et plaidoyer, Das NETTZ - Vernetzungsstelle gegen Hate Speech*
- **Jared Brown,** *Co-fondateur, Directeur Exécutif, Global Shield*
- **Ethan Chumley,** *Principal responsable de la stratégie de cybersécurité, Microsoft*
- **Styli Charalambous,** *PDG, Daily Maverick*
- **Colin Crowell,** *Conseiller Principal, Common Sense Media; Directeur général, Blue Owl Group, et ancien Vice Président pour la Politique Mondiale, Twitter*
- **Laura Edelson,** *Professeure Assistante en science de l'informatique, Northeastern University*
- **Alessandro Fabris,** *Chercheur postdoctorant, Institut Max Planck pour la sécurité et la vie privée*
- **Alberto Fernandez Gibaja,** *Chef de Programme, Digitalisation et Démocratie, International IDEA*
- **Sam Gregory,** *Directeur Exécutif, WITNESS*
- **Philipp Hacker,** *Professeur de Droit et d'Éthique sur les sociétés à l'ère du numérique, European New School of Digital Studies*
- **Sebastian Hallensleben,** *Responsable de la numérisation et de l'IA, Association VDE pour les technologies électriques, électroniques et de l'information*
- **David Evan Harris,** *Spécialiste de Politique Publique du Président de l'université, UC Berkeley*
- **Clara Helming,** *Principale responsable de politique et de plaidoyer, Algorithm Watch*
- **Kristina Irion,** *Professeure Associée, Institut du Droit de l'Information, Université d'Amsterdam*
- **Martin Kretschmer,** *Professeur en Droit de la Propriété Intellectuelle et Directeur du Centre CREATe, Université de Glasgow*
- **Amanda Leal,** *Associée, Gouvernance de l'IA, The Future Society*
- **Lucas Lasota,** *Responsable du programme juridique, Free Software Foundation Europe*
- **Michael Markovitz,** *Responsable du groupe de réflexion sur le leadership médiatique GIBS, Gordon Institute of Business Science, Université de Pretoria*
- **Jacob Mchangama,** *PDG, Justitia; Directeur du projet Avenir de la liberté d'expression de la justice*

- **Hanna Möllers,** *Secrétaire Générale adjointe, Association des journalistes allemands Deutsche-Journalisten-Verband*
- **Lucy Mwaura,** *Co-fondatrice/Directrice Générale, LWM CONSULTING*
- **Lê Nguyên Hoang,** *Fondateur, Science4all; PDG, Calicarpa*
- **Cédric O,** *Membre du Comité français sur l'IA générative, ancien Ministre d'État chargé des Affaires numériques, Conseiller co-fondateur, Mistral AI*
- **Helena Puig Larrauri,** *Co-fondatrice et responsable stratégique, Build Up*
- **Sannuta Raghu,** *Rédactrice en chef, Scroll*
- **Marietje Schaake,** *Directrice de la Politique Internationale, Centre de Politique Cyber de l'Université de Stanford; Chercheuse en Politique Internationale à l'Institut de Stanford pour l'Intelligence Artificielle centrée sur l'homme*
- **Anya Schiffrin,** *Directrice de la spécialisation Technologie, Médias et Communications, Ecole des affaires internationales et publiques, Université de Columbia*
- **Emily Tucker,** *Directrice exécutive, Centre sur la Technologie et la Vie privée, Georgetown Law*
- **Richard Wingfield,** *Directeur de Technologie et Droits de l'Homme, BSR*
- **Samuel C. Woolley,** *Professeur adjoint à l'Ecole de Journalisme et professeur adjoint, par courtoisie, à l'Ecole de l'Information, Université du Texas à Austin*
- **Antonio Zappulla,** *Directeur général, Fondation Thomson Reuters*
- **Georg Zoeller,** *Directeur, Institut de l'IA et de la Transformation*

Le Forum tient à remercier chaleureusement les experts ayant participé aux réunions thématiques sur les «Mécanismes de détection, les filigranes et la provenance du contenu» du 15 novembre 2023 et sur «La redevabilité et la responsabilité juridique» du 12 décembre, pour leur temps, leur regard éclairant et leur soutien. Nous tenons également à remercier les membres du groupe de travail qui ont participé à ces réunions.

- **Jan Beyer,** *Coordinateur de la recherche sur la démocratie numérique, Democracy Reporting International*
- **Przemyslaw Biecek,** *Université de technologie de Varsovie, projet de Partenariat mondial sur l'intelligence artificielle au sujet de la responsabilité de l'IA dans la gouvernance des médias sociaux*
- **Tapabrata Chakraborty,** *Chercheur postdoctoral, Université d'Oxford et Institut Turing*
- **Andreas Engel,** *Chercheur principal à l'Université de Heidelberg*
- **David Eyers,** *Département d'informatique, Université d'Otago*
- **Graziana Kastl-Riemann,** *avocate spécialisée dans les aspects de responsabilité liés au droit d'auteur et à la violation de la vie privée par des logiciels pilotés par l'IA*
- **Susan Leavy,** *École des études de l'information et de la communication, University College Dublin, projet de Partenariat mondial sur l'intelligence artificielle au sujet de la responsabilité de l'IA dans la gouvernance des médias sociaux*
- **Virginia Morini,** *Conseil national de la recherche italien, NIRC*
- **Judy Parnall,** *Projet Origine, Responsable des normes et de l'industrie, BBC*
- **Ugo Pagallo,** *ancien avocat et professeur de jurisprudence, Département de droit, Université de Turin ; membre du groupe d'experts créé par la Commission européenne sur la responsabilité*
- **Beatriz Saab,** *Chercheuse en démocratie numérique, Democracy Reporting International*
- **Paul D. Teal,** *École d'ingénierie et d'informatique, Université Victoria de Wellington*
- **Andrew Trotman,** *Département d'informatique, Université d'Otago*

Le Forum tient à exprimer ses chaleureux remerciements aux experts ayant soumis une contribution écrite au Groupe de Travail, partageant ainsi leurs avis éclairés.

- **Ayse Arslan,** *Alumni Oxford/Cambridge , Silicon Valley Chapter*
- **Merrin Muhammed Ashraf,** *Assistante de recherche, IT for Change, Bengaluru*
- **Paul M. Barrett,** *Directeur adjoint et chercheur chevronné, Centre NYU Stern pour le Business et les Droits humains*
- **Yoshua Bengio,** *Professeur titulaire, Université de Montréal, Fondateur et Directeur scientifique, Mila, Institut québécois d'intelligence artificielle*
- **Jennifer Brody,** *Directrice adjointe, Politique et Plaidoyer Technologie et Démocratie, Freedom House*
- **Luo Chen Jun,** *Innovateur et Plaidoyer, Sécurité en ligne et modération de contenu*
- **Laurence Dierickx,** *Chercheuse postdoctorale, Université de Bergen*
- **José van Dijck,** *Professeur de Médias et Société numérique, Université d'Utrecht*
- **Michael P. Goodyear,** *Professeur adjoint intérimaire, Faculté de droit de l'Université de New York*
- **Arijit Goswami,** *Responsable principal de l'innovation dans un grand institut de services technologiques*
- **Philipp Hacker,** *Professeur de droit et d'éthique des sociétés numériques, Nouvelle Ecole Européenne d'Études Digitales*
- **Mike Harris,** *Exonym*
- **Daniel Innerarity,** *Chaire en IA & Démocratie, Ecole de Florence de Gouvernance Transnationale, European University Institute*
- **Jordan Ogg,** *Responsable de la Politique Internationale, Ofcom*
- **Hassan Patel,** *Directeur de l'Ingénierie de la Conformité à la Politique Globale de l'IA, Fairly AI*
- **Maria Paz Canales,** *Jacqueline Rowe, Ian Barber, Global Partners Digital*
- **Ioannis Pitas,** *Professeur, Département d'informatique, Université Aristote de Thessalonique (AUTH) ; Directeur du laboratoire d'intelligence artificielle et d'analyse d'informations (AIIA)*
- **Nicholas Sewe,** *Engagement Manager, Digital Inclusion Benchmark, World Benchmarking Alliance (WBA)*
- **Nii Simmonds,** *Chercheur principal au centre GeoTech de l'Atlantic Council*
- **James Wan,** *Berserq*

Afin de renforcer la participation d'experts de différentes régions au sein du Groupe de Travail, le Forum a formalisé un partenariat avec huit organisations locales en Afrique de l'Ouest, en Afrique du Sud, au Liban et en Amérique latine :

- **Association pour la Lutte contre le Racisme,** *l'Ethnocentrisme et le Régionalisme (ALCRER), une organisation non gouvernementale défendant les droits de l'homme, la démocratie et la bonne gouvernance au Bénin*
- **Community Focus Foundation Ghana,** *une organisation de la société civile travaillant sur la gouvernance, la participation et les réformes politiques*
- **Data Leads,** *une entreprise médiatique et technologique numérique indienne*
- **Jonction,** *une organisation non gouvernementale sénégalaise travaillant sur les droits humains, la protection des données, la cybersécurité et la liberté d'expression en ligne*
- **Maharat Foundation,** *une organisation non gouvernementale pour la liberté d'expression et le développement des médias basée au Liban*
- **Media Monitoring Africa,** *une organisation sud-africaine faisant la promotion d'un journalisme éthique et juste*
- **Réseau des Professionnels de la Presse en Ligne de Côte d'Ivoire (REPPRELCI),** *la première organisation professionnelle de la presse numérique en Côte d'Ivoire*

- **Observacom (Observatoire latino-américain de la Régulation,** *des Médias et de la Convergence), un groupe de réflexion régional à but non lucratif, professionnel et indépendant, spécialisé dans la régulation et les politiques publiques liées aux médias, aux télécommunications, à l'internet et à la liberté d'expression*

Grâce à leur soutien, 61 experts ont été mobilisés et ont participé à des ateliers pour élaborer les recommandations présentées dans ce rapport.

Bénin

- **Marius Janvier Dossou-Yovo,** *Administrateur civil, Docteur d'État en droit privé, Expert en droit numérique et des médias, Coordinateur du groupe de réflexion Société de l'information : Pratiques et Gouvernance, Professeur et Chercheur*
- **Wenceslas Mahoussi,** *Maître de conférences en sciences de l'information et de la communication, École Nationale des Sciences et Techniques de l'Information et de la Communication (ENSTIC), Université d'Abomey-Calavi*
- **Yisségnon Rémy Oke,** *Ingénieur en infrastructure informatique, Directeur du Numérique, Ministère du Numérique et de la Digitalisation, et Assistant technique national pour le secteur numérique, Enabel - Agence belge de développement*
- **Toundé Seth Amon Dedehouanou,** *Ingénieur en infrastructure informatique, Ministère des Affaires numériques et de la Digitalisation*
- **Alain Codjo Cakpo,** *Ingénieur en infrastructure informatique, Chef du Service informatique, Direction Générale de l'Enseignement Supérieur*
- **Vinasétan Ratheil Esse Houndji,** *Maître de conférences en intelligence artificielle, Université d'Abomey-Calavi (UAC) ; Chef du Département d'ingénierie en logiciel, Institut de formation et de recherche en informatique, UAC ; Président, Fondation Ratheil pour une Intelligence Artificielle Responsable et Efficace*
- **Gérard Nakou,** *Ingénieur en infrastructure informatique, Gestion des systèmes d'information, municipalités de Pehunco et Kouande*
- **Hans Norbert Atacle,** *Chef de projet, ONG ALCRER, Recherche opérationnelle, Institut de Mathématiques et de Sciences Physiques, Cotonou*
- **Gervais Loko,** *Expert en gouvernance et démocratie, et Responsable de programme, ONG ALCRER*

Ghana

- **Richard Kasu,** *Analyste de politique, CFF-Ghana*
- **Martin Thompson Ntem,** *Chargé de cours et spécialiste en communications numériques, Institut de Communication en Marketing Digital, Ghana*
- **Miriam Ocloo,** *Formatrice en technologie, EM Services, Ghana*
- **Francis Kasu,** *Spécialiste en informatique de la santé publique, CFF-Ghana*
- **Abraham Dzagbletey,** *Chargé de cours et expert en marketing numérique, Institut de Communication en Marketing Digital, Ghana*
- **Princess Lovia Tetteh,** *Spécialiste de la gouvernance de l'internet et du développement des jeunes, Love Aid Foundation*
- **Albert Gharbin,** *Spécialiste en conception de communication, McGharbins Group*
- **Charles Obianim,** *Spécialiste en informatique et finance, CFF-Ghana*

Inde

- **Prof. Saima Saeed,** *Professeure associée, Centre des Médias, Droit et Gouvernance, Jamia Millia Islamia (JMI), New Delhi*
- **Prof. Taberez Neyazi,** *Professeur adjoint en communication politique et nouveaux médias, et Directeur du projet Digital Campaign Asia, Université Nationale de Singapour*

- **S.Y. Qureshi,** *Ancien Chef de la Commission électorale, Inde*
- **Arpit Chaturvedi,** *Co-fondateur et PDG, Global Policy Insights*
- **Manisha Pathak Shelat,** *Professeure en communication et plateformes numériques et stratégies, MICA, Ahmedabad, Inde*
- **Sam Daniels,** *Journaliste télévisé senior, Passionné d'IA*
- **Nisha Bhambani,** *Avocate chevronnée, Cour suprême, Inde*
- **Kavya Sukumar,** *Principale (secteurs de la technologie civique et des médias), Lightrock*

Côte d'Ivoire

- **Lassina Serme,** *Journaliste chevronné, Président, Réseau des Professionnels de la Presse en Ligne de Côte d'Ivoire (REPPRELCI), Expert en questions numériques*
- **Karim Wally,** *Journaliste ; Chargé de cours, Université Félix Houphouët-Boigny*
- **Dr Achi Harrison,** *Directeur, Laboratoire Métaverse, Université Virtuelle de Côte d'Ivoire*
- **Lucien Houedanou,** *Président, Cénacle des Journalistes Seniors de Côte d'Ivoire*
- **Mamadou Konate,** *Data Scientist, Développeur*

Amérique latine

- **Flavia Costa,** *Chercheuse adjointe, Conseil national de la recherche scientifique et technique (CONICET), Argentine*
- **Martín Becerra,** *Professeur, Universidad Nacional de Quilmes (UNQ) ; Chercheur principal, Conseil national de la recherche scientifique et technique (CONICET), Argentine*
- **Patricia Peña,** *Directrice, Datos Protegidos, Chili*
- **Juan Ortiz,** *Chercheur, Centre Berkman Klein four l'Internet et la Société, Université d'Harvard, et Fondateur de Common Ground, Argentine et États-Unis*
- **Luis Fernando García,** *Directeur exécutif, Red en Defensa De Los Derechos Digitales (R3D), Mexique*
- **Edison Lanza,** *ancien Rapporteur spécial pour la liberté d'expression, Commission interaméricaine des droits de l'homme, Uruguay*
- **Paulina Gutiérrez,** *ancienne Responsable du Programme des droits numériques chez ARTICLE 19, Mexique*
- **Maia Levy Daniel,** *Chercheuse affiliée, Centre sur la technologie et la société (CETyS), Université de San Andrés, Argentine, et ancienne Directrice de la recherche et des politiques publiques, Centro Latam Digital, Mexique*
- **Ramiro Álvarez Ugarte,** *Vice-directeur du Centre d'Études sur la Liberté d'Expression et l'Accès à l'Information (CELE), Argentine*

Liban

- **Dr. Maria Bou Zeid,** *Doyenne de la Faculté des Lettres, Université Notre Dame de Louaize (NDU)*
- **Layal Jebran,** *Experte en Technologie*
- **Zeina Bou Harb,** *Responsable de la Coopération Internationale, OGERO Telecom ; Responsable du Secrétariat Général, Forum Libanais sur la Gouvernance de l'Internet*
- **Abed Kataya,** *Responsable du Contenu Numérique, SMEX*
- **Tony Mikhael,** *Expert Juridique*
- **Dr. Marc Ibrahim,** *Directeur, Institut National des Télécommunications et de l'Informatique (INCI) ; Professeur Associé, École Supérieure des Ingénieurs de Beyrouth (ESIB), Université Saint-Joseph de Beyrouth*
- **Wael Akiki,** *Chef de Programme, Fondation Samir Kassir*
- **Layal Sakr,** *Experte Juridique, Fondatrice de Seeds for Legal Initiatives*

Sénégal

- **Ndeye Fatou Mboup,** *Jeune Experte-chercheuse, IPAR/Partenariat Mondial sur l'Intelligence Artificielle*
- **Justin Oumar Bamahossovi,** *Avocat chargé de coopération, Commission pour la Protection des Données Personnelles (CPDP) ; Chercheur en droit international de l'espace cybernétique*
- **Fana Cissé,** *Journaliste/Reporter, Média/PressAfrik*
- **Maateuw Mbaye,** *Chargé de Programme, Article 19*
- **Mouhamed Ndiaye Bocoum,** *Consultant Juridique, Commission pour la Protection des Données Personnelles (CPDP)*
- **Abdoulaye Diallo,** *Coordinateur (Information Juridique et Scientifique), Département des Droits Numériques, Réseau des Professionnels de la Presse en Ligne de Côte d'Ivoire (RADDHO)*
- **Assane Sy,** *Consultant et Formateur, UnLine Sas*
- **Emmanuel Diokh,** *Techno-Pédagogue, Formateur Juridique et Président, Internet Sans Frontières Sénégal*

Afrique du Sud

- **Dimitri Martinis,** *PDG, MCM Digital Media*
- **Unathi Malunga,** *Avocate dans l'industrie du divertissement, Consultante pour les industries créatives et Responsable du contenu*
- **Izak Minnaar,** *Expert en journalisme, médias numériques, élections et politique, Consultant et Formateur, Forum National des Éditeurs Sud-Africains (Sanef), Conseil de la Presse d'Afrique du Sud, Coalition pour le Soutien à la Radiodiffusion Publique (SOS)*
- **Kgothatso Mampa,** *Expert en Droit des Médias et Droit Commercial des Auteurs*
- **Tharin Pillay,** *Chercheur Associé, ALT Advisory*
- **Sarah Chiumbu,** *Professeure Associée, Département de Communication et Médias, Université de Johannesburg*

Le Forum tient à remercier chaleureusement les organisations partenaires qui ont soutenu le lancement de ce rapport le 28 février 2024, à savoir :

- **Centre pour une Intelligence Artificielle compatible avec l'homme** *basé à l'UC Berkeley (États-Unis)*
- **Ecole de Gouvernance Transnationale de Florence à l'Institut Universitaire Européen** *(Italie)*
- **Centre pour de Droit, Internet et Société de l'Institut pour le Développement, l'Éducation et la Recherche** *(Brésil)*
- **Research ICT Africa** *(Afrique du Sud)*
- **SciencesPo Paris Ecole de Relations Internationales, Lab Innovation Tech et Affaires Internationales** *(France)*

BIBLIOGRAPHIE

Allen, D. and Weyl, E.G. (2024). The Real Dangers of Generative AI. *Journal of Democracy*. Disponible sur : www.journalofdemocracy.org/articles/the-real-dangers-of-generative-ai/ (Consulté le 7 février 2024).

Amazon Web Services (2023). *What is an API? - API Beginner's Guide - AWS*. [online] Amazon Web Services, Inc. Disponible sur : https://aws.amazon.com/what-is/api/. (Consulté le 2 février 2024).

Amnesty International (2022). *Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – Report.* Disponible sur : www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ (Consulté le 7 février 2024).

Anderljung, M., Smith, E., O'brien, J., Soder, L., Bucknall, B., Bluemke, E., Schuett, J., Trager, R., Strahm, L. and Chowdhury, R. (2023). *Towards Publicly Accountable Frontier LLMs.* Disponible sur : https://arxiv.org/pdf/2311.14711.pdf (Consulté le 7 février 2024).

Arguedas, A., Robertson, C., Fletcher, R. and Nielsen, R. (2022). *Echo chambers, filter bubbles, and polarisation: a literature review.* Reuters Institute for the Study of Journalism. Disponible sur : https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review (Consulté le 7 février 2024).

Arisoy, E., Leipold, K. and Messan, K. (2023), The expanding role of public procurement in Africa's economic development, World Bank Blogs Disponible sur : https://blogs.worldbank.org/governance/expanding-role-public-procurement-africas-economic-development (Consulté le 9 février 2024).

Assemblée Générale des Nations Unis (2023). *Promotion et protection des droits humains dans le contexte des technologies numériques.* Disponible sur : https://digitallibrary.un.org/record/4032837?ln=fr&v=pdf (Consulté le 7 février 2024).

Awan, A.A. (2022). *A Complete Guide to Data Augmentation*. Disponible sur : www.datacamp.com/tutorial/complete-guide-data-augmentation (Consulté le 31 janvier 2024).

ASEAN (2023). ASEAN *Guide on AI Governance and Ethics.* Disponible sur : https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf (Consulté le 7 février 2024).

Associated Press (2024). *AI-powered misinformation is the world's biggest short-term threat, Davos report says.* Disponible sur : https://apnews.com/article/artificial-intelligence-davos-misinformation-disinformation-climate-change-106a1347ca9f987bf71da1f86a141968 (Consulté le 7 février 2024).

Bell, A., Stoyanovich, J. and Nov, O. (2023). *Algorithmic Transparency Playbook*. Center for Responsible AI. Disponible sur : https://dataresponsibly.github.io/algorithmic-transparency-playbook/resources/transparency_playbook_camera_ready.pdf (Consulté le 7 février 2024).

Belli, L. et al. (2022), Towards meaningful and interoperable transparency for digital platforms, Internet Governance Forum. Disponible sur : www.intgovforum.org/en/filedepot_download/57/23886 (Consulté le 7 février 2024).

Bengani, P., Stray, J., & Thorburn, L. (2022). *Blog Post: What's Right and What's Wrong with Optimizing for Engagement.* Center for Human-Compatible AI at UC Berkeley. Disponible sur : https://humancompatible.ai/news/2022/05/02/blog-post-whats-right-and-whats-wrong-with-optimizing-for-engagement/ (Consulté le 7 février 2024).

Bengio, Y. (2023). *AI and Catastrophic Risk*. Journal of Democracy. Disponible sur : www.journalofdemocracy.org/ai-and-catastrophic-risk/ (Consulté le 7 février 2024).

Bertuzzi, L. (2024). *EU countries give crucial nod to first-of-a-kind Artificial Intelligence law.* Euractiv. Disponible sur : www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/ (Consulté le 7 février 2024).

Birhane, A. et al (2021). *Multimodal datasets: misogyny, pornography, and malignant stereotypes.* Disponible sur : https://arxiv.org/abs/2110.01963 (Consulté le 7 février 2024).

Bitdefender, *What is social media impersonation?* Disponible sur : www.bitdefender.com/cyberpedia/what-is-social-media-impersonation/ (Consulté le 8 février 2024).

Bizga, A. (2020), *Blog Post: What is impersonation?,* Bitedefender. Disponible sur : www.bitdefender.com/blog/hotforsecurity/what-is-impersonation/ (Consulté le 8 février 2024).

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D. and Liang, P. (2023). *The Foundation Model Transparency Index*. Disponible sur : https://crfm.stanford.edu/fmti/fmti.pdf (Consulté le 7 février 2024).

Bontcheva, K. et al (2020), *Legislative and Regulatory Responses to Disinformation, Excerpt from the Original Report*, Broadband Commission for Sustainable Development. Disponible sur : https://en.unesco.org/sites/default/files/balanceact_legislative_en.pdf (Consulté le 8 février 2024).

Brookings. (n.d.). *Detecting AI fingerprints: A guide to watermarking and beyond*. Disponible sur : www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/ (Consulté le 31 janvier 2024).

BSA (2023). *AI Developers and Deployers: An Important Distinction.* The Software Alliance. Disponible sur : www.bsa.org/policy-filings/ai-developers-and-deployers-an-important-distinction

(Consulté le 2 février 2024).

Cambridge Consultants (2019). *Use of AI in Online Content Moderation.* Disponible sur : www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Consulté le 7 février 2024).

Chang, T.A., Arnett, C., Tu, Z. and Bergen, B.K., (2023). *When is multilinguality a curse? Language modeling for 250 high-and low-resource languages.* Disponible sur : https://doi.org/10.48550/arXiv.2311.09205 (Consulté le 7 février 2024).

Coleman, J.L. (2002). Fault and strict liability. *Risks and Wrongs*, pp.212–233. Disponible sur : https://doi.org/10.1093/acprof:oso/9780199253616.003.0012 (Consulté le 8 février 2024).

Commission Européenne (2024). *La Commission recueille actuellement des avis sur le projet de lignes directrices de la législation sur les services numériques pour l'intégrité des élections*. Disponible sur : https://digital-strategy.ec.europa.eu/fr/news/commission-gathering-views-draft-dsa-guidelines-election-integrity (Consulté le 15 février 2024).

Commission Européenne (2024). *Décision de la Commission instituant le Bureau européen de l'IA.* Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/commission-decision-establishing-european-ai-office (Consulté le 7 février 2024).

Commission Européenne (2023). *Règlement délégué sur les audits indépendants au titre de la législation sur les services numériques.* Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/delegated-regulation-independent-audits-under-digital-services-act (Consulté le 7 février 2024).

Commission Européenne (2023).*Code de conduite international pour les systèmes d'IA avancés dans le cadre du processus Hiroshima.* Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Consulté le 7 février 2024).

Commission Européenne (2022). *AI Watch: Estimating AI Investments in the European Union*. Available at: https://ai-watch.ec.europa.eu/publications/ai-watch-estimating-ai-investments-european-union_en (Consulté le 7 février 2024).

Commission Européenne (2022). *European Centre for Algorithmic Transparency*. Disponible sur : https://algorithmic-transparency.ec.europa.eu/about_en (Consulté le 7 février 2024).

Commission Européenne (2019). Lignes directrices en matière d'éthique pour une IA digne de confiance. Disponible sur : https://digital-strategy.ec.europa.eu/fr/library/ethics-guidelines-trustworthy-ai (Consulté le 15 février 2024).

Commission Européenne (2018), *A Definition of AI: Main Capabilities and Scientific Disciplines*, p.1. Disponible sur : https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (Consulté le 8 février 2024).

Conférence des Nations Unies sur le commerce et le développement. Data Protection and Privacy Legislation Worldwide. Disponible sur https://unctad.org/page/data-protection-and-privacy-legislation-worldwide (Consulté le 7 février 2024).

Conseil de l'Europe (2023). *Projet de Convention-cadre sur l'Intelligence Artificielle, les Droits de l'Homme, la Démocratie et l'Etat de Droit.* Disponible sur : https://rm.coe.int/cai-2023-28-fr-projet-de-convention-cadre/1680ae19a1 (Consulté le 7 février 2024).

Conseil de l'Europe (2023). *Lignes directrices sur la mise en œuvre responsable de systèmes d'intelligence artificielle dans le journalisme*. Disponible sur : https://rm.coe.int/cdmsi-2023-014-lignes-directrices-sur-la-mise-en-uvre-responsable-de-s/1680adb4c7 (Consulté le 7 février 2024).

Conseil européen (2024). *Législation sur l'intelligence artificielle: le Conseil et le Parlement parviennent à un accord sur les premières règles au monde en matière d'IA.* Disponible sur : https://www.consilium.europa.eu/fr/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ (Consulté le 7 février 2024).

Content Authenticity Initiative (n.d.). *How it works.* Disponible sur : https://contentauthenticity.org/how-it-works (Consulté le 7 février 2024).

CNUCED (2023), *Les Digital Government Awards de l'ONU récompensent l'excellence des services publics en ligne*. Disponible sur :

https://unctad.org/fr/news/les-digital-government-awards-de-lonu-recompensent-lexcellence-des-services-publics-en-ligne (Consulté le 7 février 2024).

c2pa.org. (n.d.). *FAQ - C2PA*. Disponible sur : https://c2pa.org/faq/ (Consulté le 31 janvier 2024).

DataLeads. *FactShala: India's Largest Media Literacy Network*. Disponible sur : https://dataleads.co.in/capacity-building/#FactShala (Consulté le 9 février 2024).

Delcker, J. (2019), *Finland's grand AI experiment*, Politico. Disponible sur : www.politico.eu/article/finland-one-percent-ai-artificial-intelligence-courses-learning-training/ (Consulté le 9 février 2024).

DeepAI. (2019). *Classifier*. Disponible sur : https://deepai.org/machine-learning-glossary-and-terms/classifier (Consulté le 2 février 2024).

Díaz, A. and Hecht-Felella, L. (2021). *Double Standards in Social Media Content Moderation,* Brennan Center for Justice. Disponible sur : www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf (Consulté le 7 février 2024).

Edelson, L., Haugen, F., and McCoy, D. (2023). *Demystifying Social Media Feeds: A Taxonomy and Transparency for Algorithmic Feed Systems Designs.* Draft manuscript.

Eisenstein, J. (2019). *Introduction to natural language processing*. Cambridge, Massachusetts: The MIT Press.

faculty.washington.edu. (n.d.). *About Data Provenance*. Disponible sur : https://faculty.washington.edu/hazeline/ProvEco/generic.html (Consulter le 1er février 2024).

Fair Trade International. About Us. Disponible sur : www.fairtrade.net/about (Consulté le 7 février 2024).

Forum sur l'information et la démocratie (2024), *Protéger les élections démocratiques par la sauvegarde de l'intégrité de l'information.* Disponible sur : https://informationdemocracy.org/wp-content/uploads/2024/02/Proteger-les-elections-democratique-par-la-souvegarde-diintegrite-de-linformation-2024.pdf (Consulté le 7 février 2024).

Forum sur l'information et la démocratie (2023). *Pluralisme de l'Information dans les Algorithmes d'Indexation et de Curation*. Disponible sur : https://informationdemocracy.org/fr/pluralisme/ (Consulté le 8 février 2024).

Forum sur l'information et la démocratie (2022). Régimes de Responsabilité pour les Réseaux Sociaux et leurs Utilisateurs. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Responsabilite-reseaux-sociaux_FR.pdf (Consulté le 8 février 2024).

Forum sur l'information et la démocratie (2020). *Pour mettre fin aux infodémies*. Disponible sur : https://informationdemocracy.org/wp-content/uploads/2023/08/ID_Infodemies_FR.pdf (Consulté le 8 février 2024).

Forum sur l'information et la démocratie. *Partenariat International sur l'Information et la Démocratie*. Disponible sur : https://informationdemocracy.org/fr/partenariat-international-information-democratie/ (Consulté le 8 février 2024).

Forum sur l'information et la démocratie. Observatoire International sur l'information et la démocratie. Disponible sur : https://informationdemocracy.org/fr/mission-2/ (Consulté le 7 février 2024).

Freckleton, I. (2020), *COVID-19: Fear, quackery, false representations and the law*, International Journal of Law and Psychiatry, Volume 72. Disponible sur : https://doi.org/10.1016/j.ijlp.2020.101611 (Consulté le 8 février 2024).

FSFE - Free Software Foundation Europe (n.d.). *What is Free Software*. Disponible sur : https://fsfe.org/freesoftware/freesoftware.en.html (Consulté le 7 février 2024).

Frontier Model Forum (2023). *What is Red Teaming?* Disponible sur : www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf (Consulté le 7 février 2024).

Fung, D.O., Brian (2023), *First on CNN: Biden campaign prepares legal fight against election deepfakes*, CNN Politics. Disponible sur : https://edition.cnn.com/2023/11/30/politics/biden-campaign-prepares-against-deepfakes/index.html (Consulté le 8 février 2024).

Funke, D. and Flamini, D. (2018), *A guide to anti-misinformation actions around the world*, Poynter. Disponible sur : www.poynter.org/ifcn/anti-misinformation-actions/ (Consulté le 8 février 2024).

Future of Life (2017). Asilomar AI Principles. Disponible sur : https://futureoflife.org/open-letter/ai-principles/ (Consulté le 7 février 2024).

Gallo, A. (2017). *A Refresher on A/B Testing*. Harvard Business Review. Disponible sur : https://hbr.org/2017/06/a-refresher-on-ab-testing (Consulté le 7 février 2024).

GeeksforGeeks, *Robustness Testing*. Disponible sur : www.geeksforgeeks.org/robustness-testing (Consulté le 9 février 2024).

Global Infodemic Management Course. Global Infodemic Management Course for Healthcare Workers. Disponible sur : https://gimch.org/ (Consulté le 9 février 2024).

Goldstein, J., Sastry, G., Musser, M., Diresta, R., Gentzel, M., Sedova, K., Adler, S., Avin, S., Bansemer, J., Bregler, C., Brundage, M., Gregory, S., Grossman, S., Herbert-Voss, A., Jernite, Y., Leibowicz, C., Leahy, C., Lin, H., Lohn, D. and Mitchell, M. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. Disponible sur : https://arxiv.org/pdf/2301.04246.pdf (Consulté le 7 février 2024).

Gordon Institute of Business and Science (2023). *Géants de la tech et journalisme : principes pour une compensation équitable.* Disponible sur : www.gibs.co.za/news-events/news/pages/big-tech-and-journalism-principles.aspx (Consulté le 7 février 2024).

Gouvernement du Canada (2023). *Loi sur l'intelligence artificielle et les données.* Disponible sur : https://ised-isde.canada.ca/site/innover-meilleur-canada/fr/loi-lintelligence-artificielle-donnees (Consulté le 7 février 2024).

GPAI (2023). *Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report,* Global Partnership on AI. Disponible sur : https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Consulté le 7 février 2024).

GPAI (2022). *Transparency Mechanisms for Social Media Recommender Algorithms: from Proposals to Actions. Report,* Global Partnership on AI. Disponible sur t: https://gpai.ai/projects/responsible-ai/transparency-mechanisms-for-social-media-recommender-algorithms.pdf (Consulté le 8 février 2024).

Grynbaum, M.M. and Mac, R. (2023). *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*. The New York Times. Disponible sur : www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (Consulté le 7 février 2024).

Halm, K.C.,Kumar, A., Segal, J. and Kalinowski, C. (2020).T*wo New California Laws Tackle Deepfake Videos in Politics and Porn. Disponible sur :* www.dwt.com/blogs/media-law-monitor/2020/02/two-new-california-laws-tackle-deepfake-videos-in (Consulté le 12 février 2024).

HCDH. Access to remedy and the technology sector: basic concepts and principles, UN B-Tech Foundational Paper. Disponible sur : www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf (Consulté le 8 février 2024).

HCDH. Taxonomy of Human Rights Risks Connected to Generative AI. Disponible sur : www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf (Consulté le 7 février 2024).

Henderson, P. (2023), *Who Is Liable When Generative AI Says Something Harmful?* Stanford University. Disponible sur :

https://hai.stanford.edu/news/who-liable-when-generative-ai-says-something-harmful (Consulté le 8 février 2024).

Hsu, T., Thompson, S.A. and Myers, S.L. (2024). *Elections and Disinformation Are Colliding Like Never Before in 2024.* The New York Times. Disponible sur : www.nytimes.com/2024/01/09/business/media/election-disinformation-2024.html (Consulté le 7 février 2024).

IBM. *What is an AI model? IBM*. Disponible sur : www.ibm.com/topics/ai-model. (Consulté le 2 février 2024).

ICPSR. *Data Management & Curation*. Disponible sur : www.icpsr.umich.edu/web/pages/datamanagement/index.html (Consulté le 8 février 2024).

Information Commissioner's Office. *A Guide to ICO Audit Artificial Intelligence (AI) Audits Contents*. Disponible sur : https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf (Consulté le 7 février 2024).

intellabs.github.io. *Knowledge Distillation - Neural Network Distiller.* Disponible sur : https://intellabs.github.io/distiller/knowledge_distillation.html (Consulté le 7 février 2024).

ISACA (2018). *Auditing Artificial Intelligence*. Disponible sur : https://ec.europa.eu/futurium/en/system/files/ged/auditing-artificial-intelligence.pdf (Consulté le 7 février 2024).

Jernite, Y; (2023). *Training Data Transparency in AI: Tools, Trends, and Policy Recommendations, Hugging Face Community Blog.* Disponible sur : https://huggingface.co/blog/yjernite/data-transparency#data-transparency-in-focus-what-is-needed (Consulté le 7 février 2024).

Ji, J., et al. (2023). *AI Alignment: A Comprehensive Survey*. arXiv (Cornell University). Disponible sur : https://doi.org/10.48550/arxiv.2310.19852. (Consulté le 2 février 2024).

Journalism Trust Initiative. Disponible sur : www.journalismtrustinitiative.org/ (Consulté le 7 février 2024).

Karathanasis, A.L., Stephanie Celis J. and Theodoros (2022). *Civil Liability for AI Systems: Comment on EU Commission's Proposals.* MIAI. Disponible sur : https://ai-regulation.com/eu-commission-proposals-on-ai-civil-liability/ (Consulté le 7 février 2024).

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. and Goldstein, T.), *A Watermark for Large Language Models.* Disponible sur : https://arxiv.org/pdf/2301.10226.pdf (Consulté le 7 février 2024).

Kovic, M., Rauchfleisch, A., Sele, M. and Caspar, C. (2018). *Digital astroturfing in politics: Definition, typology, and countermeasures. Studies in Communication Sciences*, 18(1). Disponible sur : www.hope.uzh.ch/scoms/article/view/j.scoms.2018.01.005/991 (Consulté le 7 février 2024).

Kreps, S. and Kriner, D. (2023). *How AI Threatens Democracy. Journal of Democracy*. Disponible sur : www.journalofdemocracy.org/articles/how-ai-threatens-democracy/ (Consulté le 7 février 2024).

Kretschmer, M., Kretschmer, T., Peukert, A. and Peukert, C. (2023). *The risks of risk-based AI regulation: taking liability seriously. Social Science Research Network*. Disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4622405 (Consulté le 7 février 2024).

Kulick, A. (2022), *Meta's Oversight Board and Beyond – Corporations as Interpreters and Adjudicators of International Human Rights Norms,* The Law and Practice of International Courts and Tribunals 2022, Forthcoming. Disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4226521 (Consulté le 9 février 2024).

Laux, J. (2023). *Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act*. *SSRN Electronic Journal*. Disponible sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4377481 (Consulté le 7 février 2024).

Law Insider. *deployer Definition*. [online] Disponible sur : www.lawinsider.com/dictionary/deployer (Consulté le 2 février 2024).

Leibowicz, C. (2023). *Why watermarking AI-generated content won't guarantee trust online*. MIT Technology Review. Disponible sur : www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/ (Consulté le 7 février 2024).

Lyons, K. (2020). *Facebook rolls back 'nicer' News Feed that boosted mainstream publishers*, The Verge. Disponible sur : www.theverge.com/2020/12/17/22180259/facebook-news-feed-change-post-election-publishers-misinformation (Consulté le 7 février 2024).

Magrani, E. (2019). *New perspectives on ethics and the laws of artificial intelligence.* Internet Policy Review. Disponible sur : https://policyreview.info/articles/analysis/new-perspectives-ethics-and-laws-artificial-intelligence (Consulté le 7 février 2024).

Malgieri, G. and Pasquale, F. (2024). *Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. Computer Law & Security Review*. Disponible sur : https://doi.org/10.1016/j.clsr.2023.105899 (Consulté le 7 février 2024).

Martineau, K. (2021). *What is synthetic data?* [online] IBM Research Blog. Disponible sur : https://research.ibm.com/blog/what-is-synthetic-data (Consulté le 31 janvier 2024).

Marwala, T. (2023). *Algorithm Bias — Synthetic Data Should Be Option of Last Resort When Training AI Systems.* United Nations University. Disponible sur : https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems (Consulté le 7 février 2024).

Marzotto, M. (2023), *Whistleblowers are human rights defenders. So why don't we protect them like they are?* The Signals Network. Disponible sur : https://thesignalsnetwork.org/whistleblowers-are-human-rights-defenders/ (Consulté le 9 février 2024).

Mehta, R. (2023). *Human Data Labeling for Successful AI*. [online] iMerit. Disponible sur : https://imerit.net/blog/human-data-labeling-for-successful-ai/ (Consulté le 9 février 2024).

Metaxa, D., Park, J.S., Robertson, R.E., Karahalios, K., Wilson, C., Hancock, J. and Sandvig, C. (2021). *Auditing Algorithms: Understanding Algorithmic Systems from the Outside In.* Disponible sur : https://hci.stanford.edu/publications/2021/FnT_AuditingAlgorithms.pdf (Consulté le 7 février 2024).

Miller, K and Lohn, A. (2023). *Techniques to Make Large Language Models Smaller: AN Explainers, Center for Security and Emerging Technologies.* Disponible sur : https://cset.georgetown.edu/publication/techniques-to-make-large-language-models-smaller-an-explainer/ (Accessed: 7 February 2024).

Miller, K. (2023), Introducing The Foundation Model Transparency Index, Stanford University, available at:

https://hai.stanford.edu/news/introducing-foundation-model-transparency-index (Accessed: 9 February 2024).

Miller, T., Baird, T., Littlefield, C., Kofinas, G., Chapin, F. and Redman, C. (2008). *Epistemological Pluralism: Reorganizing Interdisciplinary Epistemological Pluralism: Reorganizing Interdisciplinary Research Research. Ecology and Society*. Disponible sur : https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1034&context=unf_research (Consulté le 7 février 2024).

Milmo, D. (2023), *AI firms 'should include members of public on boards to protect society', The Guardian*. Disponible sur :

www.theguardian.com/technology/2023/dec/06/ai-firms-should-include-members-of-public-on-boards-to-protect-society (Consulté le 9 février 2024).

Mislove, A. (2023). *OSTP Blog Post: Red-Teaming Large Language Models to Identify Novel AI Risks.* The White House, Office of Science and Technology Policy. Disponible sur : www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/ (Consulté le 7 février 2024).

Mökander, J., Axente, M., Casolari, F. and Floridi, L. (2021). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, 32. Disponible sur : https://doi.org/10.1007/s11023-021-09577-4 (Consulté le 7 février 2024).

Mulder, V., Mermoud, A., Lenders, V., Tellenbach, B. (eds) *Trends in Data Protection and Encryption Technologies*. Springer, Cham. Disponible sur : https://doi.org/10.1007/978-3-031-33386-6_6 (Consulté le 7 février 2024).

Nanni, D. (2023). *Synthetic Entities: Definitions, Characteristics, and Future Perspectives*. Brass For Brain. Disponible sur : https://medium.com/brass-for-brain/synthetic-entities-definitions-characteristics-and-future-perspectives-49673f22f6fe (Consulté le 31 janvier 2024).

News/Media Alliance (2023). *White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement And Not a Fair Use.* Disponible sur : www.newsmediaalliance.org/generative-ai-white-paper/ (Consulté le 7 février 2024).

Nicholas, G. and Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis*. Center for Democracy and Technology. Disponible sur : https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Consulté le 7 février 2024).

NIST. *Computer Security Resource Center - Glossary*. csrc.nist.gov. Disponible sur : https://csrc.nist.gov/glossary/term/Red_Team (Consulté le 8 février 2024).

OCDE. N*orme d'enregistrement algorithmique transparent.* Observatory of Public Sector Innovation. Disponible sur : https://oecd-opsi.org/innovations/algorithmic-transparency-standard/ (Consulté le 7 février 2024).

OCDE. *OECD Working Party and Network of Experts on AI*. Disponible sur : https://oecd.ai/fr/network-of-experts (Consulté le 7 février 2024).

OCDE (2024), *Poids des marchés publics, Panorama des administrations publiques 2023*
Disponible sur : https://www.oecd-ilibrary.org/fr/governance/panorama-des-administrations-publiques-2023_e5bad4fb-fr (Consulté le 9 février 2024).

OCDE (2023). L'Aide Publique au Développement (APD) - *OCDE*. Disponible sur : https://www.oecd.org/fr/cad/financementpourledeveloppementdurable/normes-financement-developpement/aide-publique-au-developpement.htm (Consulté le 7 février 2024).

OCDE (2023). *Déclaration de résultat sur la Solution reposant sur deux piliers pour résoudre les défis fiscaux soulevés par la numérisation de l'économie*. Disponible sur : https://www.oecd.org/fr/fiscalite/beps/declaration-de-resultat-sur-la-solution-reposant-sur-deux-piliers-pour-resoudre-les-defis-fiscaux-souleves-par-la-numerisation-de-l-economie-juillet-2023.pdf (Consulté le 9 février 2024).

OCDE (2022). *Responsible AI licenses: a practical tool for implementing the OECD Principles for Trustworthy AI*. Disponible sur : https://oecd.ai/en/wonk/rails-licenses-trustworthy-ai (Consulté le 7 février 2024).

OCDE (2020), *Government at a Glance: Latin America and the Caribbean 2020*. Disponible sur : www.oecd.org/publications/government-at-a-glance-latin-america-and-the-caribbean-5ceda53e-en.htm (Consulté le 9 février 2024).

OCDE (2019). *Recommandation du Conseil sur l'intelligence artificielle.* Disponible sur : https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449 (Consulté le 7 février 2024).

OCDE (2019). *The OECD Artificial Intelligence (AI) Principles.* Disponible sur : https://oecd.ai/en/ai-principles. (Consulté le 7 février 2024).

OCDE (2015), *Recommandation du Conseil sur les marchés publics*. Disponible sur : https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0411 (Consulté le 9 février 2024).

Ogunleye, I. (2022). *AI's Redress Problem: Recommendations to Improve Consumer Protection from Artificial Intelligence*. Disponible sur : https://cltc.berkeley.edu/publication/cltc-white-paper-ais-redress-problem/ (Consulté le 7 février 2024).

OpenAI (2023), *Privacy policy.* Disponible sur : https://openai.com/policies/privacy-policy (Consulté le 7 février 2024).

Oracle.com (2022). *What is A/B Testing?* Disponible sur : www.oracle.com/cx/marketing/what-is-ab-testing/ (Consulté le 2 février 2024).

Organisation des Nations Unies (2001). *Déclaration universelle des droits de l'homme.* Disponible sur : https://www.un.org/fr/about-us/universal-declaration-of-human-rights (Consulté le 7 février 2024).

Organisation des Nations Unies (1948). *Déclaration universelle des droits de l'homme*. Disponible sur : https://www.un.org/fr/about-us/universal-declaration-of-human-rights (Consulté le 7 février 2024).

Organisation des Nations Unies. *La DUDH : Fondement du droit international relatif aux droits de l'homme.* Disponible sur : www.un.org/en/about-us/udhr/foundation-of-international-human-rights-law (Consulté le 8 février 2024).

Organisation des Nations Unies. *Civil society*. Disponible sur : www.un.org/en/civil-society/page/about-us (Consulté le 8 février 2024).

OSCE (2021), *Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual*. Disponible sur : www.osce.org/files/f/documents/8/f/510332_1.pdf (Consulté le 7 février 2024)

Oversight Board. Disponible sur : www.oversightboard.com/ (Consulté le 7 février 2024).

Parlement Européen (2022). *Auditing the quality of datasets used in algorithmic decision-making systems.* Disponible sur : www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf (Consulté le 7 février 2024).

Parlement Européen (2020), *Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission sur un régime de responsabilité civile pour l'intelligence artificielle*. Disponible sur : https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_FR.html (Consulté le 8 février 2024).

Partnership on AI. PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action. Disponible sur : https://syntheticmedia.partnershiponai.org/ (Consulté le 7 février 2024).

PNUD. Improving Procurement Transparency. Disponible sur : www.undp.org/asia-pacific/fairbiz/improving-procurement-transparency (Consulté le 7 février 2024).

Project Origin. *Project Origin.* Disponible sur : www.originproject.info/about (Consulté le 7 février 2024).

Public Citizen (2023), *Comment to FEC: A.I.-Generated Political Deepfakes Are 'Fraudulent Misrepresentation.'* Disponible sur : www.citizen.org/article/comment-to-fec-a-i-generated-political-deepfakes-are-fraudulent-misrepresentation/ (Consulté le 7 février 2024).

Règlement (UE) 2022/2065 du Parlement européen et du Conseil du 19 octobre 2022 relatif à un marché unique des services numériques et modifiant la directive 2000/31/CE (règlement sur les services numériques). Disponible sur : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A32022R2065 (Consulté le 7 février 2024).

Responsible AI Licenses. Disponible sur : www.licenses.ai/ (Consulté le 7 février 2024).

Robertson, D. (2023). *When 'red-teaming' AI isn't enough.* Politico. Disponible sur : www.politico.com/newsletters/digital-future-daily/2023/10/25/when-red-teaming-ai-isnt-enough-00123577 (Consulté le 7 février 2024).

Routley, N. (2023). *What is generative AI? An AI explains*. World Economic Forum. Disponible sur : www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/ (Consulté le 9 février 2024).

RSF (2023), RSF et 16 organisations partenaires présentent la Charte de Paris sur l'IA et le journalisme. Disponible sur : https://rsf.org/fr/rsf-et-16-organisations-partenaires-pr%C3%A9sentent-la-charte-de-paris-sur-l-ia-et-le-journalisme (Consulté le 7 février 2024).

RSF (2023), *« Projet Spinoza » : RSF et l'Alliance de la presse d'information générale partenaires pour développer un outil d'intelligence artificielle pour les journalistes*. Disponible sur : https://rsf.org/fr/projet-spinoza-rsf-et-l-alliance-de-la-presse-d-information-g%C3%A9n%C3%A9rale-partenaires-pour-d%C3%A9velopper (Consulté le 9 février 2024).

Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N. and Allen, P. (2019). *The Risk of Racial Bias in Hate Speech Detection.* Association for Computational Linguistics. Disponible sur : https://aclanthology.org/P19-1163.pdf (Consulté le 7 février 2024).

Sarno, I. (2023). *What Is a Large Language Model?* Disponible sur : https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model (Consulté le 8 février 2024).

Schakowky, J. (2021), A Bill to provide incentives for and protect whistleblowers under the authority of the Federal Trade Commission, and for other purposes. Disponible sur : https://schakowsky.house.gov/sites/evo-subsites/schakowsky-evo.house.gov/files/SCHAKO_082_xml.pdf (Consulté le 7 février 2024).

Schwartz, M. (2008). The Trolls Among Us. *The New York Times*. Disponible sur : www.nytimes.com/2008/08/03/magazine/03trolls-t.html (Consulté le 7 février 2024).

Shavit, Y. et al. (2023). *Practices for Governing Agentic AI Systems.* Disponible sur : https://openai.com/research/practices-for-governing-agentic-ai-systems (Consulté le 5 février 2024).

Shorten, C. and Khoshgoftaar, T.M. (2019). *A survey on Image Data Augmentation for Deep Learning.* Journal of Big Data, [online] 6(1). Disponible sur : :https://doi.org/10.1186/s40537-019-0197-0. (Consulté le 14 février 2024).

Sneha Solanki (2024). *What is criminal liability? Definition and resources for defense attorneys*. Thomson Reuters Law Blog. Disponible sur : https://legal.thomsonreuters.com/blog/what-is-criminal-liability/ (Consulté le 9 février 2024).

Sijbrandij, S. (2023), *AI weights are not open «source»*, Open Core Ventures. Disponible sur : https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/ (Consulté le 9 février 2024).

Silberg, J. and Manyika J. (2019). *Notes from the AI frontier: Tackling bias in AI (and humans)*. McKinsey Global Institute. Disponible sur : www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans (Consulté le 7 février 2024).

Simchon, A., Edwards, M. and Lewandowsky, S. (2024). *The persuasive effects of political microtargeting in the age of generative AI.* Disponible sur : https://academic.oup.com/pnasnexus/article/3/2/pgae035/7591134 (Consulté le 7 février 2024).

South African Competition Commission (2023), *Final Terms of Reference (ToR) for the Media and Digital Platforms Market Inquiry*, Government Gazette No. 49309. Disponible sur : www.gov.za/sites/default/files/gcis_document/202309/49309gon3880.pdf (Consulté le 7 février 2024).

Stanford University Human-Centered Artificial Intelligence (2023). *Artificial Intelligence Index Report 2023 Introduction to the AI Index Report 2023*. Disponible sur : https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Consulté le 7 février 2024).

Stepanov, A. and Gupta, A. (2021). *Reducing Political Content in News Feed.* Meta. Disponible sur : https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/ (Consulté le 7 février 2024).

Stray, J., Iyer, R., and Puig Larrauri, H. (2023). T*he Algorithmic Management of Polarization and Violence on Social Media.* Knight First Amendment Institute at Columbia University. Disponible sur : https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media (Consulté le 7 février 2024).

Tanenbaum,W., Song, K. and Malek, L. (2022), *Theories of AI liability: It's still about the human element*, Reuters. Disponible sur : www.reuters.com/legal/litigation/theories-ai-liability-its-still-about-human-element-2022-09-20/ (Consulté le 7 février 2024).

The Bletchley Declaration by Countries Attending the AI Safety Summit. 1-2 November 2023. Disponible sur : www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 (Connsulté le 7 février 2024).

The Global Partnership on Artificial Intelligence (2020). *About - GPAI*. Disponible sur : https://gpai.ai/about/ (Consulté le 7 février 2024).

The National Congress of Brazil (2023). *Bill 2338/2023 'Brazilian Artificial Intelligence Act'*. Disponible sur : https://mcusercontent.com/af97527c75cf28e5d17467eaa/files/248d109f-eeef-7496-4df1-12d29affb522/PL_23382023_Senado_ENG_VF.pdf (Consulté le 7 février 2024).

The United States (1996). *Communications Decency Act, Section 230 of the Telecommunications Act of 1996. 47 U.S.C. § 230*. Disponible sur : www.govinfo.gov/content/pkg/USCODE-2021-title47/pdf/USCODE-2021-title47-chap5-subchapII-partI-sec230.pdf (Consulté le 8 février 2024).

The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Disponible sur : www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Consulté le 7 février 2024).

Turing (n.d.). *Fine-Tuning LLMs: Overview, Methods & Best Practices*. Disponible sur : www.turing.com/resources/finetuning-large-language-models#what-is-fine-tuning (Consulté le 9 février 2024).

Union Européenne (2019), *Liability for artificial intelligence and other emerging digital technologies,* disponible sur : https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en (Consulté le 7 février 2024).

UK Government (2023), *Emerging Processed for AI safety.* Disponible sur : https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (Consulté le 8 février 2024).

UNESCO (2023). *Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence.* Disponible sur : www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence (Consulté le 7 février 2024).

UNESCO (2023), *Multilevel and Meaningful Transparency in Algorithmic Systems: Developing Concrete Criteria to Guide Institutional and Legal Reforms.* Disponible sur : www.unesco.org/en/articles/multilevel-and-meaningful-transparency-algorithmic-systems-developing-concrete-criteria-guide (Consulté le 7 février 2024).

UNESCO (2022). *Déclaration de Windhoek sur l'intelligence artificielle en Afrique australe Windhoek.* Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000383197_fre (Consulté le 7 février 2024).

UNESCO (2022), *Enseigner l'intelligence artificielle au primaire et au secondaire : une cartographie des programmes validés par les gouvernements*. Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000380602 (Consulté le 8 février 2024).

UNESCO (2021). *Recommandation sur l'éthique de l'intelligence artificielle.* Disponible sur : https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Consulté le 7 février 2024).

UNESCO (2020). *Le journalisme, « les fausses nouvelles » (fake news) et désinformation : un manuel pour l'enseignement et la formation du journalisme*. Disponible sur : https://fr.unesco.org/fightfakenews (Consulté le 8 février 2024).

United Nations Secretary-General's AI Advisory Body (2023). *Interim Report: Governing AI for Humanity*. Disponible sur : www.un.org/techenvoy/sites/www.un.org.techenvoy/files/ai_advisory_body_interim_report.pdf (Consulté le 7 février 2024).

US Department of Labor. Retaliation. Disponible sur : www.whistleblowers.gov/know_your_rights (Consulté le 9 février 2024).

US Department of Labor (n.d.), *How to Create an Anti-Retaliation Program.* Disponible sur : www.whistleblowers.gov/antiretaliation (Consulté le 9 février 2024).

US Office of Cyber and Infrastructure Analysis (2018). Social Media Bots Overview. Disponible sur : https://niccs.cisa.gov/sites/default/files/documents/pdf/ncsam_socialmediabotsoverview_508.pdf?trackDocs=ncsam_socialmediabotsoverview_508.pdf (Consulté le 8 février 2024).

Webb, A. (2019), *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*, Public Affairs New York.

Web Foundation et al; (2018), Universal Service and Access Funds: An Untapped Resource to Close the Gender Digital Divide. Disponible sur : https://webfoundation.org/docs/2018/03/Using-USAFs-to-Close-the-Gender-Digital-Divide-in-Africa.pdf (Consulté le 9 février 2024).

Wendy Hall, D. and Pesenti, J (2017). *Growing the Artificial Intelligence Industry in the UK*. Disponible sur : https://assets.publishing.service.gov.uk/media/5a824465e5274a2e87dc2079/Growing_the_artificial_intelligence_industry_in_the_UK.pdf (Consulté le 7 février 2024).

Wheeler, T. (2023). *The three challenges of AI regulation.* Brookings. Disponible sur : www.brookings.edu/articles/the-three-challenges-of-ai-regulation/ (Consulté le 7 février 2024).

World Economic Forum (2024). *The Global Risks Report 2024*. Disponible sur : www.weforum.org/publications/global-risks-report-2024/ (Consulté le 7 février 2024).

Yeung, K. (2019, *Responsibility and AI*, Council of Europe study. Disponible sur : https://rm.coe.int/responsability-and-ai-en/168097d9c5 (Consulté le 8 février 2024).

Yu, N. et al (2022). *Artificial Fingerprinting for Generative Models*: *Rooting Deepfake Attribution in Training Data.* Disponible sur : https://arxiv.org/abs/2007.08457 (Consulté le 7 février 2024).

Yun Chee, F (2024). *EU lawmakers ratify political deal on artificial intelligence rules.* Reuters. Disponible sur : www.reuters.com/technology/eu-lawmakers-back-political-deal-artificial-intelligence-rules-2024-02-13/ (Consulté le 13 février 2024).

Zewe, A. (2022), *In machine learning, synthetic data can offer real performance improvements, MIT News.* Disponible sur : https://news.mit.edu/2022/synthetic-data-ai-improvements-1103 (Consulté le 7 février 2024).

Zhang, H. et al. (2023). *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models.* Disponible sur : https://arxiv.org/abs/2311.04378 (Consulté le 15 février 2024).

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M. and Lin, M. (n.d.). *A Recipe for Watermarking Diffusion Models.* Disponible sur : https://arxiv.org/pdf/2303.10137.pdf (Consulté le 7 février 2024).

Zhou, J., Zhang, Y., Luo, Q., Parker, A.G. and Munmun De Choudhury (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.* Disponible sur : https://jiaweizhou.me/assets/chi23_ai_misinfo.pdf (Consulté le 7 février 2024).

Soutenu par


FORD
FOUNDATION


Luminate
Building stronger societies


NED
NATIONAL
ENDOWMENT
FOR
DEMOCRACY
SUPPORTING FREEDOM AROUND THE WORLD

Forum Information et Démocratie
CS 90247
75083 Paris Cedex 02, France
contact@informationdemocracy.org
https://informationdemocracy.org

Forum Information & Démocratie