CDMSI(2025)15
3 décembre 2025

Note d'orientation sur les implications de l'intelligence artificielle générative sur la liberté d'expression

Adopté par le Comité directeur du Conseil de l'Europe sur les médias et la société de l'information (CDMSI) lors de sa 28e réunion plénière, 3-5 décembre 2025 (sous réserve de modifications éditoriales)

Introduction - Définition et champ d'application

1. Les États membres du Conseil de l'Europe se sont engagés à garantir à toute personne relevant de leur juridiction les droits et libertés consacrés par la Convention de sauvegarde des droits de l'homme et des libertés fondamentales (STE n° 5, « la Convention »). Cet engagement demeure valable tout au long du processus continu de progrès technologique et de transformation numérique que connaissent les sociétés européennes.

2. L'article 10 de la Convention consacre le droit à la liberté d'expression, qui « comprend la liberté d'opinion et la liberté de recevoir ou de communiquer des informations ou des idées ». La Cour européenne des droits de l'homme (« la Cour ») rappelle, dans son abondante jurisprudence, que la liberté d'expression, en ligne et hors ligne, est l'un des fondements de la société démocratique ainsi que l'une des conditions essentielles de son progrès et de l'épanouissement de chacun[1]. L'exercice réel et effectif de ce droit, qui ne dépend pas uniquement du devoir de l'État de ne pas interférer de manière négative, nécessite également des mesures positives de protection, même dans la sphère des relations entre les individus.

3. Des instruments récents du Conseil de l'Europe ont souligné que l'évolution rapide de l'environnement numérique, ainsi que le développement des systèmes d'intelligence artificielle (IA), peuvent favoriser le progrès individuel et collectif, l'inclusion sociale et l'innovation, tout en présentant des risques susceptibles de compromettre divers droits fondamentaux et de fragiliser des valeurs démocratiques, tels que le droit à la liberté d'expression[2].

4. La Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit (STCE N. 225) du 2024 dispose que les activités menées au sein du cycle de vie des systèmes d'intelligence artificielle doivent être pleinement compatibles avec les droits humains, la démocratie et l'État de droit, tout en étant propices au progrès technologique et à l'innovation[3].

5. Le domaine de l'IA connaît actuellement un essor important, notamment dans sa forme générative. Largement accessible et simple d'utilisation pour des finalités diverses, l'IA générative attire différentes catégories d'utilisateurs, notamment des particuliers, des entreprises privées et des institutions publiques.

6. Ici, le terme « IA générative », désigne un système d'IA composite pouvant produire de nouveaux contenus ou de nouvelles sorties qui avoisinent les productions humaines, sur la base de motifs identifiés dans les données d'entraînement. Dotés de degrés variables d'autonomie et d'interaction avec l'utilisateur, les systèmes d'IA générative peuvent produire du texte, des images, des sons, des vidéos, des actions ou une combinaison de ces éléments, ou encore transformer des contenus selon différentes modalités et formats.

7. Les systèmes d'IA générative facilitent la création de contenus et permettent des nouvelles formes de communication et d'expression. Ils contribuent ainsi au développement d'applications utiles et enrichissantes qui permettent de diffuser informations et connaissances grâce à la génération automatisée de contenus. Malheureusement, ces systèmes peuvent également servir à des fins de persuasion, de manipulation ou d'activités malveillantes, et

reproduire, voire accentuer, les inégalités existantes au sein de la société, ce qui peut compromettant l'exercice effectif de la liberté d'expression et d'autres droits et libertés.

8. Les systèmes d'IA générative permettent de nouvelles formes d'hyperpersonnalisation avancée de l'expérience utilisateur caractérisées par des contenus uniques à chaque utilisateur. Ces propriétés sont susceptibles d'exercer un impact significatif sur l'écosystème de l'information, en accentuant la fragmentation de la diffusion des contenus informationnels au profit d'une « audience au singulier », où chaque utilisateur interagit de manière solitaire, isolée et automatisée avec des contenus informationnels façonnés en fonction de son profil singulier. Cette transition d'une audience plurielle à une « audience au singulier » porte atteinte à l'espace informationnel partagé et pluraliste qui est fondamental à la démocratie.

9. En raison de l'ample adoption de l'IA générative pour le recueil, le partage d'informations, et la communication d'idées ou la formation des opinions, cette technologie a la possibilité d'influer de manière significative sur les différentes formes d'opinion et d'expression, et de peser sur le débat public, la diffusion du savoir, la création de contenus et leur distribution.

10. L'IA générative se caractérise également par une évolution continue, tant sur le plan technologique que dans ses applications concrètes. Un tel progrès, en particulier lorsqu'il est rapide, est susceptible de renforcer les apports positifs de cette technologie pour la liberté d'expression, mais également d'en accentuer les risques.

11. Des études confirment les craintes que suscitent le manque de transparence, de qualité, d'exactitude, de reproductibilité, de fiabilité, d'équité et de factualité des contenus générés par l'IA, questions que le présent document entend examiner sous l'angle du droit à la liberté d'expression. En effet, toutes les dimensions de la liberté d'expression peuvent être affectées par l'IA générative, tant à l'échelle individuelle qu'à celle de la société, et ce à court, moyen et long terme.

12. L'objectif de la présente note d'orientation est triple :

i. **jeter les bases d'une compréhension commune** des implications de l'IA générative sur le droit à la liberté d'expression, en forgeant une terminologie partagée, un cadre d'analyse et de référence communs qui facilitent le dialogue entre l'ensemble des parties prenantes ;

ii. **identifier systématiquement les implications structurelles** des systèmes basés sur l'IA générative pour la liberté d'expression ; et

iii. **offrir un ensemble concret de mesures actionnables** à l'intention des décideurs politiques, principalement les États membres, mais aussi les fournisseurs de technologies, la société civile et d'autres parties prenantes concernées, au sein d'un cycle de gouvernance agile qui cibles les implications structurelles de cette technologie conformément à la Convention.

13. Aux fins de la présente note d'orientation, et afin d'analyser les implications sur la liberté d'expression des systèmes basés sur l'IA générative, leur cycle de vie est considéré comme étant composé de trois couches technologiques principales : la technologie de fondation (« couche fondamentale »), la phase de développement d'outils (« couche outil ») ainsi que la conception et l'optimisation d'applications (« couche produit »).

14. La note d'orientation se concentre uniquement sur les implications de l'IA générative sur la liberté d'expression. Compte tenu des interactions complexes de la liberté d'expression avec d'autres libertés et droits fondamentaux, les aspects connexes ne sont abordés que de manière générale ou incidente. Si les questions relatives, par exemple, à la vie privée (article 8 de la Convention), à l'interdiction de la discrimination (article 14 de la Convention), aux droits des enfants et des personnes vulnérables, à la propriété intellectuelle (y compris les droits d'auteur) et à l'impact sur l'environnement sont, dans certains cas d'usage, interdépendantes et indissociables, elles ne sont pas l'objet de cette note et ne font pas l'objet d'un traitement approfondi. Les implications et les interactions de ces droits mériteraient une analyse et un

rapport plus étayés, mais en l'absence de directives supplémentaires, il serait bon de leur accorder une attention lors de la mise en œuvre des actions et recommandations formulées dans la présente note.

15. Étant donné que les implications de l'IA générative sont nombreuses, encore insuffisamment étudiées et en constante évolution, ce document n'a pas pour objet de fournir un aperçu exhaustif de l'ensemble des domaines potentiellement concernés.

16. Le présent document est structuré en quatre sections. La première présente les principales caractéristiques technologiques de l'IA générative et de son cycle de vie en constante évolution, désigné sous le terme de « stack technologique de l'IA générative » (ou *Tech Stack*). La deuxième examine l'article 10 de la Convention dans le contexte en question. La troisième propose une analyse des implications structurelles de l'utilisation de l'IA générative sur la liberté d'expression, à partir de cas d'usages connus. Enfin, la quatrième section formule des orientations visant à maximiser les bénéfices de cette technologie tout en réduisant les risques associés.

17. La note d'orientation s'appuie sur les documents et standards existants du Conseil de l'Europe, en particulier la Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'Etat de droit, ainsi que, *inter alia*, les Recommandations du Comité des Ministres CM/Rec(2018)2 sur les rôles et responsabilités des intermédiaires de l'internet, CM/Rec(2020)1 sur les impacts des systèmes algorithmiques sur les droits de l'homme, CM/Rec(2022)4 sur la promotion d'un environnement favorable à un journalisme de qualité à l'ère numérique, CM/Rec(2022)11 sur les principes de gouvernance des médias et de la communication, CM/Rec(2022)13 sur les effets des technologies numériques sur la liberté d'expression, et les Lignes directrices sur la mise en œuvre responsable des systèmes d'intelligence artificielle dans le journalisme, adoptées par le Comité directeur sur les médias et la société de l'information (CDMSI) en 2023.

18. La note d'orientation s'appuie sur les idées, les connaissances et les expériences d'un large éventail d'acteurs qui ont contribué à son élaboration, notamment les membres du Comité d'experts du Conseil de l'Europe sur les implications de l'IA générative pour la liberté d'expression (MSI-AI).

SECTION 1 - LA STACK TECHNOLOGIQUE DE L'IA GENERATIVE : LA COUCHE FONDAMENTALE, LA COUCHE OUTIL ET LA COUCHE PRODUIT

19. **La stack technologique de l'IA générative :** A travers la stack technologique de l'IA générative sont décrites des étapes fondamentales de son cycle de vie et des processus qui sont utilisés pour concevoir, déployer et maintenir les systèmes et applications qui en découlent. Elle se compose de trois couches principales : la couche technologique fondamentale, la couche des outils et la couche des produits. Ces couches mettent en jeu différents processus technologiques, certaines conditions et facteurs clé du développement cycle de vie des systèmes d'IA générative, tels que la puissance de calcul, les données et les compétences, ainsi que différents acteurs économiques et parties prenantes, susceptibles d'avoir une implication sur la qualité, l'exactitude, la fiabilité, et sur la présence, plus ou moins marquée, de biais dans les contenus générés par l'IA.

20. **Implications à chaque couche de la stack :** A chaque couche de la stack technologique de l'IA générative peuvent émerger des implications distinctes pour la liberté d'expression. La cartographie actuelle de ces couches technologiques permet d'identifier les avantages et les risques spécifiques qui apparaissent tout au long du cycle de vie de l'IA générative, tels qu'ils peuvent être compris au moment de l'élaboration du présent document, compte tenu de l'évolution rapide de cette technologie et de ses applications (voir figure 1). Les avantages et les risques associés à certains usages seront examinés à la Section 3, afin d'illustrer en quoi une approche guidée par une compréhension de la stack technologique est essentielle pour identifier et analyser les implications du cycle de vie de l'IA générative sur la liberté d'expression.

21. **La couche fondamentale :** La première couche correspond à celle des modèles de fondation, là où s'opère la phase initiale d'entraînement du modèle d'IA de base. Ces modèles

sont développés au moyen de processus d'apprentissage machine qui nécessitent une capacité de calcul conséquente ainsi qu'un volume considérable de données d'entraînement (voir figure 1, étapes 1 à 3).

22. **Les données d'entraînement :** Les résultats générés par le modèle de base sont liés aux motifs extraits des données d'entraînement. Il est crucial de veiller à l'adoption de bonnes pratiques pour constituer des données d'entraînement représentatives, ainsi qu'à leur étiquetage et à leur prétraitement appropriés (voir figure 1, étapes 1 et 2), afin de réduire au minimum le risque de biais dans les modèles d'IA générative. Des exemples documentés de productions biaisées en raison du genre[4], de la race[5] ou d'autres facteurs, montrent qu'il existe des problèmes dans les données intégrées dans les corpus d'entraînement ou post-entrainement, ils proviennent parfois d'informations de mauvaise qualité, voire des informations erronées[6]. Tout contenu généré qui est biaisé ou trompeur parce que les données sont de qualité médiocre ou non représentatives, peut avoir de graves répercussions sur la liberté d'expression, en particulier sur le droit de recevoir des informations et de former et détenir des opinions. La qualité et l'évaluation des données d'entraînement sont déterminantes pour assurer un premier niveau de contrôle des biais.

23. **La diversité linguistique et culturelle des données d'entraînement :** Le manque de diversité linguistique et culturelle dans les données d'entraînement est un problème important qui se pose au niveau de la couche fondamentale et qui a des répercussions sur la représentation des différentes cultures et contextes de ces langues. Bien que des améliorations soient en cours dans ce domaine, la langue anglaise reste surreprésentée dans les données d'entraînement. Ce déséquilibre linguistique a une incidence directe sur la liberté d'expression des utilisateurs qui parlent des langues minoritaires ou faiblement dotées en ressources numériques,[7] et qui sont également moins susceptibles d'avoir un accès équitable à des informations de qualité et d'en bénéficier, dans leur langue maternelle, par le biais d'applications fondées sur l'IA générative.

24. **La couche outils :** La deuxième couche transforme les modèles de fondation en outils automatisant certaines tâches, par exemple en transformant un grand modèle de langage généraliste en un système qui produit des questions-réponses. Plusieurs problèmes distincts se posent à ce stade en matière de liberté d'expression, notamment lorsqu'il faut adapter les modèles de base pour créer des outils interactifs et assistants IA afin qu'ils suivent les instructions des utilisateurs en exécutant des tâches précises, telles que la synthèse, la traduction ou la reformulation (voir figure 1, étape 4). À cette étape, le contenu généré par le modèle de fondation est ultérieurement adapté, par diverses techniques, à des préférences humaines spécifiques (voir figure 1, étape 5). Des systèmes de filtrages ou politiques de modération des contenus générés interviennent aussi par exemple pour de refuser l'accès à des instructions sur la fabrication d'armes ou éviter les réponses discriminatoires (voir figure 1, étape 6).

25. **Les risques de sycophanterie :** Un risque spécifique apparaît au niveau de la couche outil, notamment lorsque les modèles de base sont adaptés de manière à privilégier la satisfaction de l'utilisateur, ainsi que la personnalisation de l'expérience, au détriment de l'exactitude des faits ou du respect du pluralisme des points de vue (voir figure 1, étape 5). Des études ont montré par exemple que les contenus produits par l'IA générative tendent à refléter les convictions de l'utilisateur, à adopter des positions politiques similaires ou à faire preuve de complaisance en flattant ou en adaptant les réponses dans le but de prolonger l'interaction ou de favoriser une conversation plus conviviale. Cette tendance trompeuse, qualifiée de « sycophanterie », résulte des processus d'apprentissage par renforcement mis en œuvre au sein de la couche outil[8] (voir figure 1, étape 5). Elle conduit à la génération de contenus hyper-personnalisé, voire persuasifs ou trompeurs, qui renforcent les comportements, croyances et préjugés de l'utilisateur. Les outils et applications d'IA générative fonctionnant comme des « chambres d'écho » ou « bulles de filtre » sont susceptibles de porter atteinte au droit de chacun de se forger une opinion, ainsi qu'au droit d'accéder à une information exacte, diversifiée et fondée sur la pluralité des idées[9]. L'exercice effectif du droit à la liberté d'expression (inclut le droit à la liberté d'opinion) nécessite en effet un accès à des contenus pluralistes issus de sources variées[10].

26. **Les risques liés au filtrage et aux garde-fous :** Le recours à des filtres et des garde-fous permet aux outils d'IA générative de mettre en œuvre des formes de filtrage qui peuvent s'apparenter à une modération de contenu (voir figure 1, étape 6). Toutefois, si ces filtres et

mécanismes ne sont pas conçus de manière appropriée et proportionnée, et conformément aux normes en matière de liberté d'expression[11], ils risquent de devenir des formes d'influence indue, de manipulation ou, dans le pire des cas, de censure. Ils peuvent également affecter la portée et l'intégrité de l'information journalistique et des médias dans le nouvel environnement de recherche et d'information intermédiée par l'IA. De plus, une modération inadéquate ou négligée des contenus peut même favoriser la prolifération de la discrimination et des discours haineux[12].

27. **La couche produit :** Au sein de la troisième et dernière couche de la stack technologique de l'IA générative, les outils fondés sur l'IA générative sont personnalisés et optimisés pour devenir des produits interagissant avec les utilisateurs. L'accent est mis ici sur la conception de produits et services basés sur l'IA générative, tels qu'applications interactives, *chatbots* ou agents IA[13] et qui assistent l'utilisateur dans la recherche, la collecte d'informations, la génération de contenus et autant d'automatisation de tâches et orchestrations de processus. À cette étape, divers types de techniques d'optimisation et de personnalisation sont employés. Ceux-ci peuvent inclure la contextualisation de réponses afin de rechercher et d'utiliser des sources d'information données pour générer des réponses plus fiables (appelée génération augmentée par récupération de données, ou RAG)[14], des fonctionnalités axées sur le design et la conception de produits, telles que les suggestions de prompt et les fonctions mémoire dans les *chatbots*, ou encore des systèmes d'IA générative plus complexes, tels que des agents d'IA capables d'exécuter plusieurs tâches en parallèle et de manière plus autonome (voir figure 1, étapes 7 à 10).

28. **Les risques liés à la conception et au design de l'expérience utilisateur :** Les techniques qui permettent de créer des applications sur mesure pour les utilisateurs finaux suscitent des inquiétudes quant à la manière dont les produits basés sur l'IA générative et leur conception de l'expérience utilisateur peuvent influencer la liberté d'expression de l'utilisateur, intentionnellement ou non, indirectement ou non. Il a été démontré que ces techniques peuvent avoir des effets néfastes sur l'interaction, tels que la persuasion personnalisée, le renforcement des stéréotypes ou la contrainte au passage à l'acte. Par exemple, plusieurs produits d'IA générative intègrent des fonctions de mémorisation permettant de conserver des informations issues d'interactions passées, ce qui révèle des éléments relatifs à l'identité et aux préférences des utilisateurs, lesquels sont ensuite utilisés pour influencer les interactions futures (voir figure 1, étapes 8, 9 ou 10). Si cette fonctionnalité favorise des conversations plus personnalisées et contextuellement pertinentes, donnant l'impression d'interactions naturelles et continues, elle suscite également des préoccupations en matière d'influence trompeuse, d'autonomie cognitive[15], d'anthropomorphisme, de biais, de respect de la vie privée, de non-discrimination et de vulnérabilité. Cela est particulièrement problématique si les utilisateurs sont traités différemment sur la base d'attributs mémorisés tels que le genre ou l'identité, qui sont déduits ou supposés sur la base des interactions passées des utilisateurs avec une application d'IA générative, telle qu'un *chatbot* conversationnel[16]. Des préoccupations encore plus fortes apparaissent dans le contexte des LLM multimodaux et des agents d'IA, lorsque les informations mémorisées lors d'interactions passées sont utilisées pour simuler un comportement humain[17] et prédire les prochaines étapes, intentions ou même les achats des utilisateurs, avec une précision et une capacité d'adaptation inédites[18].

29. **Les agents d'IA et les effets cumulatifs dus à l'évolution de la stack technologique de l'IA générative :** Les effets produits à travers les différentes phases du cycle de vie tendent à se cumuler et à se renforcer mutuellement, en particulier au sein de système composites comme les agents IA. Par exemple, si les processus d'apprentissage par renforcement mis en œuvre au niveau de la couche outil (voir étape 5) promeuvent la rechercher de satisfaction de l'utilisateur, ces effets peut être accentués par le fait que la couche produit conserve les conversations et les données personnelles des utilisateurs (voir étape 10), afin de mieux déduire ce que les utilisateurs sont susceptibles d'apprécier pour permettre une interaction plaisante, fluide et interactive. Cet effet est encore renforcé par l'utilisation de techniques (telles que l'affinage par apprentissage par renforcement et l'optimisation) et par l'utilisation d'agents IA capables d'effectuer plusieurs tâches à la fois et d'automatiser ces étapes à travers différents niveaux de la stack technologique. Garantir la qualité, l'exactitude, la fiabilité, la reproductibilité, la transparence, l'exactitude factuelle et l'équité de modèles et systèmes d'IA générative suppose un suivi technologique rigoureux et à toutes les étapes de leur cycle de vie, allant de la qualité et la représentativité des corpus de données utilisés pour former les modèles de base (couche

fondamentale) ; aux affinages post-entraînement et adaptations effectuées par les développeurs pour définir les paramètres de contrôle sur les contenus générés (couche outil) ; jusqu'aux ajustements dynamiques destinés à personnaliser les produits et services grâce à l'interaction avec les utilisateurs (couche produit).

30. **Les dynamiques de marché de l'IA générative et l'importance des données des l'utilisateurs finaux :** Les dynamiques de marché à l'œuvre dans la stack technologique de l'IA générative peuvent avoir des répercussions sur la liberté d'expression. Ces effets sont amplifiés lorsque les fournisseurs sont présents de manière intégrée à travers l'ensemble de la stack. Si les étapes computationnelles de pré-entrainement relèvent principalement de la puissance de calcul et des coûts associés à l'inférence des modèles et l'exécution systèmes, c'est avant tout la disponibilité de données de haute qualité, notamment celles qui sont issues des utilisateurs finaux, qui permet l'amélioration continue des produits et services fondés sur l'IA générative. Les données des utilisateurs finaux (comme les données personnelles, l'historique des requêtes, les données comportementales d'interaction) sont autant de facteurs critiques permettant l'améliorer des modèles de base, des outils et des produits d'IA générative. Les grandes entreprises technologiques disposent d'un accès privilégié à ces données, et peuvent ainsi optimiser leurs produits, renforcer leur attractivité auprès des utilisateurs, et générer en retour un flux accru de données[19], ce qui alimente un cercle vertueux dont elles sont les principales bénéficiaires. C'est à ce stade que la concentration verticale du marché apparaît le plus clairement.

31. **La capture des données et les barrières à l'entrée** : Cette concentration verticale du marché[20] crée des barrières à l'entrée particulièrement élevées pour les nouveaux acteurs et renforce le rôle de gardien d'accès exercé par un nombre restreint d'entreprises en position dominante. Elle réduit aussi de manière significative la transparence et la capacité des acteurs externes (même lorsqu'il s'agit d'expert en technologie ou régulateurs) à observer ce qui se déroule au niveau de la couche produit, et limite ainsi la possibilité d'identifier des risques majeurs pour la liberté d'expression et l'État de droit. S'il convient de reconnaître l'existence de plusieurs initiatives ayant permis la mise en place d'outils de suivi des incidents et de taxonomies des risques[21], force est de constater qu'un écart significatif demeure en ce qui concerne la surveillance des restrictions indues à la liberté d'expression. Cette situation appelle à la mise en œuvre de mécanismes de supervision et de transparence plus robustes, en particulier au niveau de la couche produit.

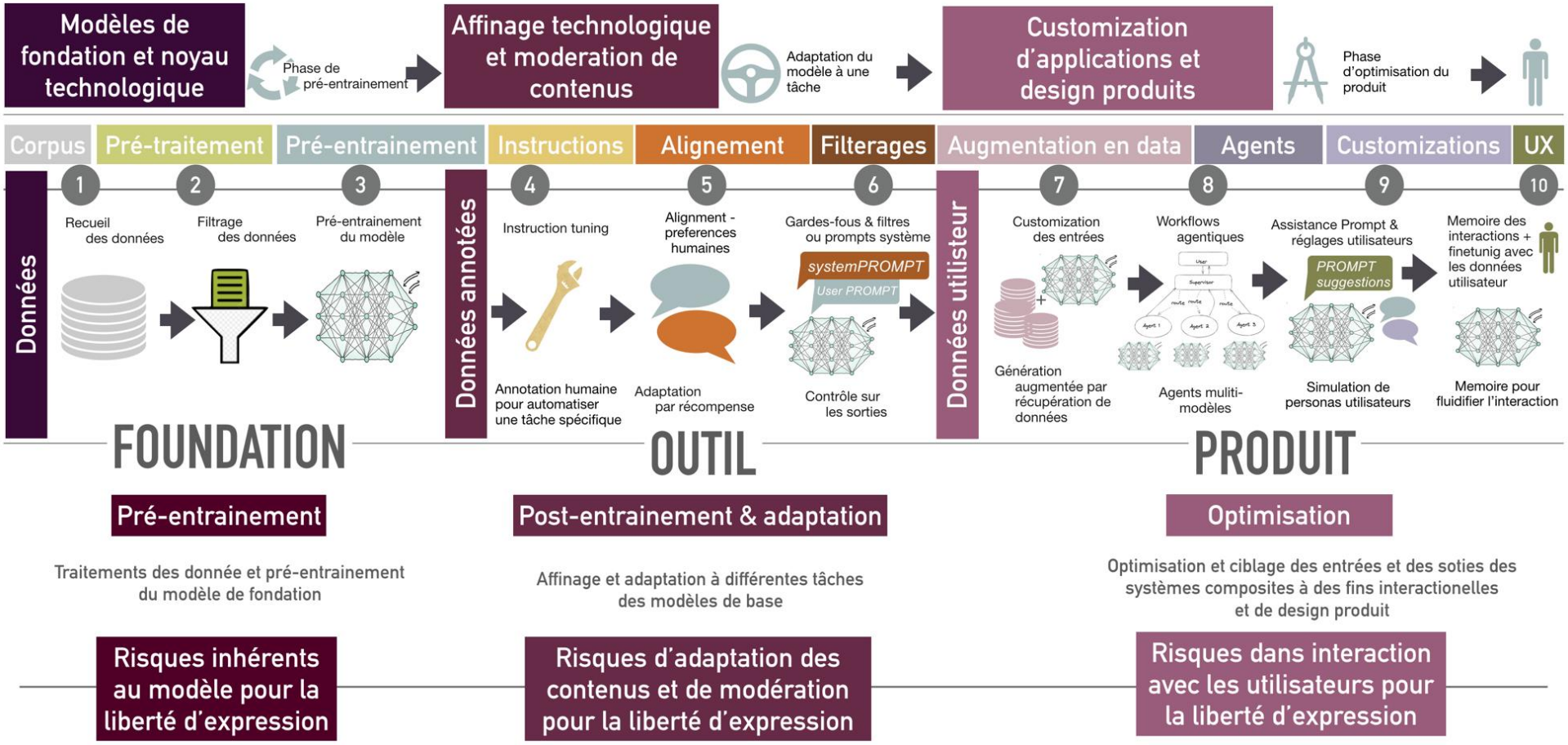


Figure 1 : La stack technologique d'IA générative de la collecte de données à l'interaction avec l'utilisateur final - Une approche stratifiée de risques liés à la liberté d'expression et consciente des différents acteurs qui interviennent pendant le cycle de vie technologique face.

SECTION 2 - LA LIBERTÉ D'EXPRESSION IA GÉNÉRATIVE : ENJEUX ET USAGES

32. La présente section examine la manière dont l'article 10 de la Convention européenne des droits de l'homme, les standards du Conseil de l'Europe Council pertinents, et la jurisprudence de la Cour européenne des droits de l'homme orientent la protection de la liberté d'expression dans le contexte de l'utilisation de l'IA générative et tout au long de son cycle de vie. Cette section met l'accent sur les obligations positives des États membres de créer un environnement propice à la liberté d'expression et de favoriser le pluralisme du débat public et la liberté des médias. Elle évalue l'utilisation de systèmes à base d'IA générative à travers le prisme de l'article 10 et propose des critères pour évaluer l'expression assistée par l'IA et sa protection éventuelle en tant qu'expression humaine.

33. Conformément au paragraphe 2 de l'article 10 de la Convention, l'exercice de la liberté d'expression comporte des devoirs et des responsabilités et peut faire l'objet d'exceptions, qui doivent être prévues par la loi, poursuivre l'un des buts légitimes au sens de l'article 10 et être nécessaires dans une société démocratique[22].

34. Afin de créer et de garantir des conditions favorables à la liberté d'expression au sens de l'article 10, les États membres doivent non seulement respecter leurs obligations négatives de non-ingérence, mais aussi s'acquitter d'un ensemble d'obligations positives. Certaines de ces obligations s'appliquent également aux systèmes d'IA générative, telles que la promotion d'un débat public ouvert, pluraliste et inclusif, ainsi que la lutte contre les contenus préjudiciables et illégaux, tout en garantissant la légitimité, la proportionnalité, la nécessité et la transparence. Les États ont un rôle à jouer pour instaurer un environnement propice à un journalisme de qualité, y compris à une information de qualité destinée au public. Cela devrait s'appliquer même dans un contexte d'évolution technologique rapide, qui peut se révéler particulièrement perturbateur pour la profession et pour son rôle démocratique[23].

35. Le Conseil de l'Europe a souvent examiné les responsabilités des acteurs privés en matière de droits de l'homme et de libertés fondamentales[24], Dans le contexte de systèmes, et selon la Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme, les acteurs du secteur privé doivent faire preuve de diligence raisonnable en matière de droits humains afin de « s'assurer qu'ils ne commettent pas de violations des droits de l'homme ou qu'ils n'y contribuent pas et que leurs actions » et d'« éviter de favoriser ou de perpétuer la discrimination tout au long du cycle de vie de leur système »[25]. Ce principe a également été reconnu dans les Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme, adoptés à l'unanimité par le Conseil des droits de l'homme des Nations Unies en 2011, qui fournissent un cadre aux gouvernements et aux entreprises pour identifier, prévenir, atténuer et réparer les violations des droits humains liées aux activités des entreprises.

36. La Cour a souligné que la démocratie repose fondamentalement sur la liberté d'expression[26]. Protégée par l'article 10, le droit à la liberté d'expression comprend la « liberté des opinions et de recevoir et de communiquer des informations et des idées sans ingérence et sans considération de frontières ». Elle s'applique non seulement aux « informations » ou « idées » accueillies favorablement ou considérées comme inoffensives ou indifférentes, mais aussi à celles qui heurtent, choquent ou inquiètent. De cette manière, la liberté d'expression permet un débat public vigoureux, qui constitue une autre condition préalable à une société démocratique caractérisée par le pluralisme, la tolérance et l'esprit d'ouverture.

37. La jurisprudence de la Cour reconnaît en outre que les médias et les journalistes qui ont un comportement éthique et responsable bénéficient d'une protection renforcée au titre de l'article 10, en raison de leur rôle fondamental dans la diffusion d'informations et de points de vue diversifiés et pluralistes. C'est sur cette base que les individus peuvent se forger des opinions éclairées, les exprimer librement et participer au débat public[27].

38. En ce qui concerne la liberté d'expression et Internet, la Cour a noté à plusieurs reprises que les activités expressives générées par les utilisateurs sur Internet constituent une plateforme sans précédent pour l'exercice de la liberté d'expression.[28] En outre, la jurisprudence de la Cour

sur la question du droit à l'oubli est pertinente pour la liberté d'expression à l'ère des nouvelles technologies[29].

39. Si la Cour n'a pas encore statué sur des affaires d'IA générative, sa jurisprudence abondante au titre de l'article 10 fournit des principes directeurs qui permettent d'évaluer les implications potentielles de l'IA générative sur le droit à la liberté d'expression[30]. En outre, un aspect essentiel du droit à la liberté d'expression à l'ère de l'IA tient au fait que la Convention doit être considérée comme « un instrument vivant, interprété à la lumière des conditions actuelles »[31].

40. La question de savoir si, en vertu de l'article 10 de la Convention, l'expression assistée par IA générative devrait bénéficier de la même protection et est soumise aux mêmes limitations que l'expression humaine[32] continue de faire débat. La note d'orientation propose que les critères suivants devraient être pris en considérations lors d'une telle évaluation[33] :

i. la question de savoir si l'expression est produite sous l'impulsion d'un individu ou de manière partiellement ou totalement automatisée, voire dans un cadre autonome par l'intermédiaire d'un agent IA[34] ;

ii. les choix technologiques et de conception à chaque couche de la stack technologique de l'IA générative et la logique qui les sous-tend, ce qui inclut l'analyse de la manière dont le modèle est construit, entraîné, optimisé, évalué et déployé, ainsi que l'intention et l'impact de ces décisions de conception sur la liberté d'expression (voir les mesures de transparence à la section 4) ;

iii. le contenu de ce qui est communiqué par l'expression humaine, étant donné que le résultat produit avec le concours de l'IA générative, qu'il soit généré ou assisté, repose sur l'invite de l'utilisateur ainsi que sur des expressions préexistantes extraites, présentes ou mémorisées dans les données d'entraînement[35] ; et

iv. la relation entre l'apport humain et le résultat structuré ou assisté par l'IA générative, compte tenu de la mesure dans laquelle ce résultat reflète, transforme ou s'écarte de l'intention initiale de l'utilisateur (voir Implications structurelles 2 et 4).

SECTION 3 – LES IMPLICATIONS STRUCTURELLES DE L'IA GÉNÉRATIVE SUR LA LIBERTÉ D'EXPRESSION

41. Les implications de l'IA générative sur la liberté d'expression, mises en évidence dans la présente note d'orientation, constituent un instantané qui ne saurait refléter la trajectoire future du développement technologique de l'IA générative. Ces implications peuvent également varier selon la manière dont cette technologie est conçue et mise à disposition des utilisateurs finaux, ainsi du contexte dans lequel elle est déployée, y compris les circonstances sociales, politiques, économiques et autres.

42. Ce document se concentre sur les implications qui, tant au niveau individuel que sociétal, présentent un caractère structurel dans la mesure où elles : a) impactent les fondements de la liberté d'expression, b) sont ancrées dans les fondements du fonctionnement concret de la technologie, et c) ne sont pas susceptibles de changer rapidement. Les observations présentées reposent cependant sur les usages actuels, mais leur pertinence et leur impact sont appelés à évoluer à mesure que la technologie de l'IA générative progresse.

43. Comme pour d'autres technologies, les avantages et les risques tiennent non seulement à la conception et aux limites structurelles de la technologie, mais aussi à la manière dont elle est utilisée. Les produits et services d'IA générative peuvent accroître l'efficacité des utilisateurs et offrir des fonctionnalités auparavant inaccessibles. Parallèlement, l'IA générative et son potentiel multimodal – texte, vidéo et images – peuvent être exploités à des fins malveillantes et causer des dommages importants tant à l'échelle individuelle que sociétaux, le contenu qu'elle produit devenant plus persuasif[36], plus facilement reproductible et adaptable à des groupes sociaux spécifiques pour en accroître l'impact[37].

44. En raison des risques associés à la conception des systèmes et à leur utilisation, les entreprises qui développent et déploient des applications d'IA générative mettent en œuvre divers mécanismes pour contrer ces risques (voir filtres et garde-fous, Figure 1 étapes 5 et 6), tels que des politiques d'alignement et de modération du contenu[38]. Si ces politiques présentent des avantages évidents, elles comportent également le risque d'une modération trop large ou insuffisante, ce qui, dans les deux cas, est susceptible de nuire à la liberté d'expression.

45. Les effets négatifs sur la liberté d'expression sont particulièrement probables lorsque les pratiques d'encadrement et de modération sont automatisés, manquent de supervision humaine et ne tiennent pas compte de la diversité linguistique ou des nuances contextuelles (par exemple, dans les cas d'expression artistique, de parodie ou de satire). La Note d'orientation du Conseil de l'Europe sur la modération de contenu énonce des principes fondamentaux susceptibles de guider une approche de la modération fondée sur les droits humains, tels que le respect des droits humains par défaut, la transparence, l'existence d'un cadre juridique et opérationnel clair, la proportionnalité, l'instauration de garanties contre une application excessive ou trop prudente des obligations de modération, ainsi que la mise en place de mécanismes de recours indépendants.

46. Cette note d'orientation, qui correspond au stade actuel de développement et d'adoption des systèmes d'IA générative, recense six domaines qui ont des implications et incidences structurelles et fondamentales sur la liberté d'expression :

i. **Expression et de l'accès au contenu :** Les systèmes fondés sur l'IA générative peuvent faciliter la diffusion de contenus, accroître les possibilités de compréhension grâce à l'adaptation interactive de ces contenus, et offrir de nouvelles modalités de partage et de réception d'opinions et d'idées (voir Implication structurelle 1, section 3.1.).

ii. **Diversité et normalisation de l'expression humaine :** Alors que les applications d'IA peuvent rendre possible de nouveau formats d'expression individuelle, elles peuvent aussi avoir un impact sur la diversité de l'expression humaine en normalisant les contenus et en réduisant la singularité de l'expression individuelle à grande échelle (voir Implication structurelle 2, section 3.2.).

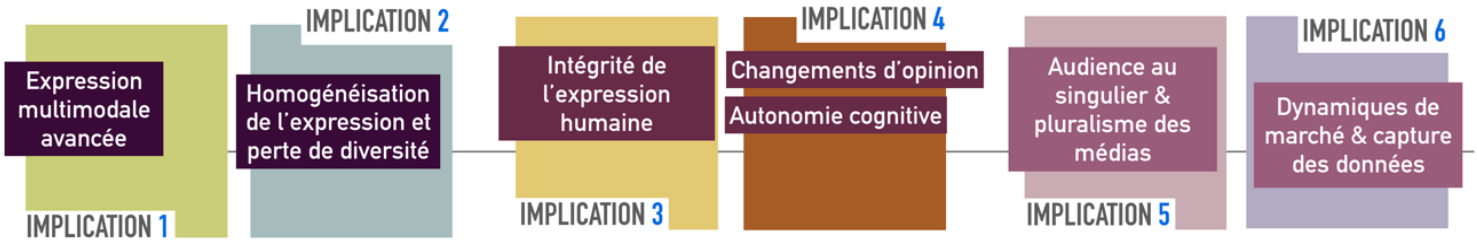
iii. **Intégrité de l'expression humaine et son attribution :** Les systèmes fondés sur l'IA générative produisent des contenus en agrégeant des phrases probables de manière statistique, sans nécessairement avoir accès à des sources variées et sans attribution explicite. Même lorsque les systèmes d'IA générative sont enrichis par des bases de données sélectionnées, les sources sont souvent mal attribuées, ce qui peut causer un préjudice à la réputation de leurs auteurs originaux (individus ou organisations). Cela rend plus difficile la tâche des utilisateurs qui cherchent à identifier et vérifier correctement la source de l'information (voir Implication structurelle 3, section 3.3.).

iv. **Capacité d'agir et formation de l'opinion :** Si les systèmes fondés sur l'IA générative peuvent à la fois fusionner différentes sources d'information et dissocier un contenu informatif de son contexte et de son auteur d'origine, leur capacité démontrée à transmettre de manière persuasive du contenu peut compromettre l'autodétermination et la faculté de se forger des opinions indépendantes. Cela peut être détournée pour influencer à grande échelle les opinions d'individus et de groupes dans l'espace public, et entraîner des changements d'opinion automatisés, qu'il s'agisse d'une persuasion admissible ou d'une manipulation inacceptable[39]. La capacité de former et d'exprimer sa propre opinion est ainsi menacée, ce qui affecte en définitive l'autonomie cognitive et, plus largement, l'intégrité de l'espace informationnel (voir Implication structurelle 4, section 3.4.).

v. **Pluralisme des médias et de l'information :** Les applications fondées sur l'IA générative peuvent remodeler l'espace public de l'information d'une manière qui remet en question le pluralisme des médias et de l'information, c'est-à-dire la diversité des opinions, des points de vue et des sources qui reflètent la pluralité des sociétés[40]. À mesure que les services pilotés par l'IA générative deviennent des points d'accès privilégiés à l'information, de nouveaux intermédiaires s'interposent entre les médias et

le public. La conception et la modération des contenus de ces applications ont ainsi un impact direct sur la visibilité et la viabilité du journalisme, ainsi que sur son rôle démocratique. Cet impact est particulièrement préoccupant lorsque les sources sont dissociées ou incorrectement attribuées, et lorsque les organisations médiatiques ne sont pas équitablement rémunérées pour l'utilisation de leurs contenus dans les tâches d'entraînement ou d'adaptation des modèles (voir Implication structurelle 5, section 3.5.).

vi. **Dynamiques de marchés** : Différents niveaux de concentration de marché peuvent être observés à différentes couches de la stack technologique de l'IA générative. Ces dynamiques, particulièrement marquées aux niveaux outil et produit, peuvent exercer un effet restrictif sur l'exercice du droit à la liberté d'expression. Motivé par des intérêts économiques ou des raisons idéologiques, le contrôle exercé sur cette infrastructure technologique, et ses processus, peut conduire à une modération insuffisante ou excessive, ainsi qu'à la diffusion de contenus filtrés, censurés, ou sélectionnés et générés par des systèmes automatisés (voir Implication structurelle 6, section 3.6.).



3.1. Implication structurelle 1 : Amélioration de l'expression individuelle et de l'accès aux contenus

47. **Facilité d'utilisation et interactivité :** Les apports de l'IA générative à la liberté d'expression tiennent à la fois à la simplicité d'utilisation de ces applications et aux modalités interactives engageantes qu'elles offrent, au service de l'expression individuelle. Fonctionnant selon un principe d'interaction en langage naturel, dans lequel l'utilisateur formule une question, une demande ou des instructions, et l'application génère du contenu sous divers formats, ces technologies aident les individus à accéder à l'information, à exprimer et articuler du contenu, des informations et des idées. L'impact est d'autant plus fort que ces technologies sont adoptées par un large public de plus en plus rapidement[41]. À la différence des moteurs de recherche traditionnels, qui extraient et présentent des informations existantes, les applications fondées sur l'IA générative produisent et agrègent statistiquement de nouveaux contenus à partir des requêtes d'utilisateurs. Cet avantage reste toutefois tributaire de la possibilité pour chacun d'accéder à l'IA générative dans sa propre langue, ainsi que d'autres conditions sociétales plus larges (accès à l'internet, fracture numérique, etc.).

48. **Meilleure accessibilité au contenu multimodal :** En tant que technologie qui permet la production, l'adaptation et l'accessibilité de contenus et d'informations, l'IA générative peut contribuer à lever les obstacles liés au savoir-faire technique, à la langue, au style et aux formats, rendant ainsi des questions complexes plus accessibles à un public plus large. Cette technologie peut particulièrement être bénéfique aux personnes handicapées[42], d'autant que les fonctionnalités multimodales, par exemple la transcription de la parole en texte ou la conversion d'images en parole, viennent renforcer l'accessibilité. En définitive, elle peut contribuer plus largement à l'exercice du droit des individus de recevoir et de communiquer des informations et des idées.

49. **Amélioration des formes d'expression humaine :** L'IA générative abaisse les barrières d'accès aux secteurs créatifs et peut encourager et soutenir la création artistique ainsi que sa diffusion multimodale, notamment la production de parodies et de contenus qui repoussent les limites sociétales et favorisent l'autoréflexion, contribuant ainsi au pluralisme et à l'inclusion. Elle peut ainsi promouvoir la diversité de l'expression humaine et inciter un plus grand nombre de personnes à participer aux débats publics sur des questions d'intérêt général, ou encore

permettre une diffusion plus large de contenus qui, autrement, resteraient limités à une seule forme (par exemple le texte).

50. **Personnalisation des contenus :** Les outils d'IA générative peuvent améliorer l'accès au contenu et aux informations d'intérêt public en produisant des messages ciblés et personnalisés, ce qui contribue à une meilleure information des citoyens. Dans le cadre du débat public, les *chatbots* ou agents conversationnels alimentés par l'IA générative peuvent fournir aux électeurs des informations personnalisées sur l'actualité, les questions politiques et autres, sous forme de texte, de voix ou d'autres formats. Une telle interaction peut favoriser une meilleure compréhension des enjeux publics, améliorer l'accès à des contenus informatifs et faciliter la formation de l'opinion publique, à condition de contrôler les usages abusifs.

51. **Nouveaux outils pour les médias, le journalisme et les industries créatives :** Les systèmes basés sur 'IA générative peuvent améliorer l'accès à l'information et être bénéfique pour les institutions et les médias, qui jouent un rôle important pour la démocratie et la liberté d'expression, en leur permettant de développer de nouvelles façons d'informer et de dialoguer avec leur public. Les outils d'IA générative d'agrégation, d'analyse, de contextualisation et de synthèse de contenus peuvent appuyer les enquêtes journalistiques, la recherche de contenus et la diffusion médiatique. Les systèmes d'IA générative peuvent aider le secteur des médias et les industries créatives à créer, adapter et distribuer des contenus, à condition que les droits d'auteur et les droits de propriété intellectuelle, ainsi que le droit à la vie privée, à la réputation et d'autres droits susceptibles d'être affectés dans ce contexte, soient clairement établis et respectés. En ce qui concerne plus particulièrement les journalistes et les médias, l'IA générative offre la possibilité d'améliorer l'accès et la recherche de sources et d'informations de manière plus générale, et de présenter leurs reportages de manière plus accessible.[43]

3.2. Implication structurelle 2 : Diversité et standardisation de l'expression

52. **Perte de diversité sociétale et homogénéisation de l'expression humaine à grande échelle** : Les systèmes d'IA générative reposent sur des modèles statistiques et probabilistes. De ce fait, ils produisent intrinsèquement des résultats qui reflètent les données d'entraînement les plus représentées et ce de manière opaque. Ainsi, ils peuvent généraliser certaines idées par le biais de processus d'affinage et adaptation avancés et de garde-fous (voir figure 1, risques liés à la modération de contenu, étapes 4, 5 et 6). Si leur impact n'est pas immédiatement perceptible au niveau individuel, leur utilisation à grande échelle peut entraîner des conséquences sociétales importantes et des incidences sur la diversité de l'expression humaine ainsi que sur la qualité des informations et contenus disponibles. L'une de ces conséquences est l'homogénéisation de l'expression humaine à grande échelle, où des voix uniques et diverses risquent d'être éclipsées par des contenus répétitifs ou statistiquement standardisés. Il s'agit d'un défi croissant non seulement pour la liberté d'expression et l'accès des individus à une information diversifiée, mais aussi pour la société au sens large, où les différentes langues et cultures, ou encore l'expertise et la réputation de ceux qui contribuent à la diversité du débat public (journalistes, experts, individus et communautés) risquent d'être standardisées ou diluées. Les effets cumulatifs d'une telle homogénéisation à grande échelle peuvent menacer la liberté d'expression et le pluralisme[44].

53. **Standardisation de l'expression individuelle :** Sur le plan individuel, cette standardisation est préoccupante car elle peut entraîner un appauvrissement de la diversité de l'expression dans la sphère privée, en réduisant la variété des points de vue au lieu de l'élargir[45]. Des études empiriques menées en situation réel mettent en évidence une perte de diversité de l'expression humaine causée par une standardisation des contenus écrits ou visuels à grande échelle. Concrètement, des participants invités à produire des idées (par exemple, dans des exercices de conception de produits) au moyen d'un outil d'IA générative montrent une amélioration notable de la qualité perçue des idées produites par chaque participant, tandis que l'ensemble de la population enregistre une réduction significative de la diversité lexicale et conceptuelle des formulations avec, dans certains cas, une baisse de 41 % de la diversité[46]. Ces tests empiriques suggèrent que l'effet à grande échelle de l'utilisation d'assistants d'IA générative entraîne une homogénéisation linguistique[47] et culturelle[48] ainsi qu'une standardisation de l'expression des utilisateurs et des idées qu'ils véhiculent à l'échelle individuelle[49]. Cela pourrait

potentiellement conduire à une perte à long terme des capacités cognitives nécessaires pour effectuer les tâches qui ont été automatisées.[50]. Ces effets de standardisation ne se limitent pas à la création automatisée de contenu écrit ou oral, des effets similaires se produisent également dans le domaine visuel[51]. Il importe en particulier d'identifier et d'éviter les impacts potentiellement discriminatoires sur les minorités linguistiques, culturelles et sociales, d'autant qu'ils peuvent résulter de données d'entraînement biaisées ou de choix de conception exclusifs.

54. **Manque de représentativité dans les corpus de données d'entrainement :** Même si des acteurs de l'IA générative (privés, universitaires, etc.) ont commencé à développer des pratiques communes en matière de collecte, de filtrage et de prétraitement des données d'entrainement, l'observation des systèmes en fonctionnement et de leurs résultats met en évidence une réalité persistante, à savoir qu'aucun jeu de données d'entraînement ne représente pleinement la diversité des catégories existantes. Il apparaît donc nécessaire d'étudier et d'améliorer l'impact des pratiques de collecte des données sur la liberté d'expression. En particulier, la diversité linguistique et culturelle doit être considérée comme une condition préalable à la représentativité et à l'inclusion. Elle doit donc être prise en compte dès la phase de conception des modèles[52] afin de garantir, par exemple, que les langues, ainsi que les minorités et les cultures faiblement dotées en ressources, ne soient pas exclues et puissent également bénéficier de l'IA générative dans le contexte de la liberté d'expression.

3.3. Implication structurelle 3 : Intégrité de l'expression humaine et de son attribution

55. **Hallucination :** Le mécanisme de fonctionnement qui repose sur la prédiction du prochain mot et caractère le plus probable entre fréquemment en conflit avec celui de véhiculer des faits établis dans les sorties des systèmes d'IA générative. Nombre d'études montrent que les systèmes d'IA générative produisent de manière récurrente des réponses inexactes, voire inventent des sources en générant statistiquement du contenu[53]. Si des améliorations technologiques ont bien été apportées pour corriger les inexactitudes de la recherche assistée par une IA générative, le phénomène d'« hallucination » constitue une menace directe pour le droit de chacun à accéder à une information fiable, qui est un élément fondamental de l'exercice de la liberté d'expression. Ce risque dépasse le cadre individuel. En effet, à l'échelle de la société, la généralisation de l'usage des produits d'IA générative peut favoriser la diffusion massive d'informations erronées en aggravant la désinformation[54], nuire à la confiance et, de manière plus générale, compromettre les systèmes d'information.

56. **Absence ou brouillage des sources d'information :** En ce qui concerne l'exactitude des informations, les applications pilotées par l'IA générative sont fondamentalement différentes des moteurs de recherche car elles créent du contenu en agrégeant statistiquement des séquences linguistiques. Elles offrent ainsi une nouvelle expérience de consommation de contenu qui n'a pas de sources identifiables, voire des sources souvent inexactes, et ce même au sein des systèmes de recherche augmentée de type « RAG » (voir étape 7 de la figure 1)[55]. Cette configuration diffère de l'environnement informationnel tels qu'il était avant l'intermédiation de l'IA, lorsqu'il reposait en effet sur des contenus humains distincts tels que des articles ou des vidéos associées à leur auteur. Le passage à la recherche et à l'information intermédiée par l'IA générative présente un risque pour le droit d'accéder à l'information et de se former une opinion, car il peut diminuer ou supprimer la possibilité ou la capacité des personnes à évaluer le contenu en fonction de ses sources.

57. **Dissociation de l'auteur :** Les résultats de l'IA générative peuvent dissocier une œuvre de son auteur, cela peut entraîner une perte de contrôle de l'auteur sur son expression et porter atteinte au droit de communiquer des informations, tout en érodant la confiance dans l'écosystème informationnel. Cela peut aussi diluer la qualité de l'expression ou de l'information d'un auteur et nuire à la réputation de l'auteur, par exemple en produisant des résumés superficiels mettant l'accent de manière trompeuse sur certains éléments. Les auteurs ont mis en garde contre le risque que les machines s'approprient leur style personnel ou leurs caractéristiques, ce qui affaiblirait et diluerait la valeur et l'originalité de leur travail et de leur identité[56]. En outre, lorsque les outils d'IA générative fournissent des informations incorrectes et les attribuent à des sources crédibles, les utilisateurs sont plus susceptibles de considérer ces

informations incorrectes comme crédibles, ce qui, à long terme, érode la confiance dans des informations exactes et vérifiées[57].

58. **Appropriation d'image et *deep fakes* :** L'usage détourné des outils d'IA générative permet l'appropriation de l'image d'autrui, la contrefaçon, l'usurpation d'identité ainsi que la banalisation des contenus falsifiés de type *deep fakes* (hypertrucages). La création et la diffusion publique de contenus falsifiés ou contrefaits visant à imiter une personne sont le plus souvent réalisées sans consentement, et peuvent s'apparenter à de véritables actes de falsification numérique. Les *deep fakes* et autres productions audiovisuelles hyperréalistes générées à l'aide de l'IA générative constituent un risque élevé pour le débat public et l'intégrité de l'information, en particulier dans le contexte des processus électoraux[58]. Le potentiel de manipulation du contenu, notamment à des fins de désinformation,[59] ou d'usurpation de l'identité de candidats, de journalistes et de personnalités publiques, est un risque important associé aux outils d'IA générative. Les *deep fakes* sont souvent utilisés pour porter atteinte à l'image des personnes visées, notamment pour affaiblir la crédibilité des femmes qui s'expriment dans l'espace public[60].

59. **Appropriation vocale et clonage vocal :** Dans le domaine du clonage vocal, le risque est particulièrement élevé pour les personnes dont les voix sont largement accessibles en ligne ou présentes dans divers répertoires publics[61]. Des cas de clonage abusif, voire potentiellement illicite, de voix appartenant à des professionnels du secteur vocal, suivis de leur commercialisation sans autorisation, ont déjà été constatés[62]. Cette situation soulève de vives préoccupations quant au droit des personnes, dont les prestations vocales sont accessibles aux entreprises développant des systèmes d'IA générative, de contrôler l'usage de leur voix et d'en garantir l'authenticité. Les incidents liés au clonage vocal révèlent une dilution croissante de l'expression personnelle dans un environnement saturé de déclarations fausses ou générées automatiquement. Contrairement à d'autres formes de simulation multimodale, le clonage vocal présente des risques spécifiques et accrus en matière de vie privée, de sécurité et d'intégrité personnelle, et de fraude.

60. **Imitation de la personnalité des individus :** Les systèmes d'IA générative et l'évolution vers des agents IA renforcent les préoccupations liées à l'appropriation qui peut être faite de l'expression des individus. Les avancées technologiques permettent d'accéder facilement à des ressources permettant aux systèmes d'IA générative d'imiter les comportements, les attitudes, la ressemblance et la personnalité de personnes réelles en utilisant très peu de données à caractère personnel[63]. Cela ouvre ainsi à de nouvelles possibilités de tromperie et de dilution de la liberté d'expression, et notamment à la perte d'attribution et d'autonomie dans l'expression individuelle. Les agents IA peuvent reproduire de manière réaliste la personnalité, les gestes, la voix et les attributs d'un individu, et refléter ses valeurs et préférences afin d'agir ou d'accomplir des tâches numériques en son nom, avec ou sans son consentement explicite.

61. **Délégitimation ou utilisation abusive d'acteurs influents ou de sources importantes :** L'IA générative peut également être utilisée abusivement pour détourner ou usurper des voix éminentes, telles que celles de journalistes, de défenseurs des droits humains ou de responsables politiques, par exemple en générant et diffusant à grande échelle à leur sujet des informations fausses, inexactes ou trompeuses, ou en les imitant (connu comme « *spoofing* »). Ces pratiques affectent déjà les organisations médiatiques et la diffusion d'informations provenant de sources fiables[64]. Le brouillage des lignes entre contenu authentique et synthétique, exact et falsifié, peut aggraver les campagnes de dénigrement et le harcèlement, qui visent en particulier les voix féminines[65]. Toutes ces évolutions peuvent également avoir un effet dissuasif ou paralysant sur l'expression de voix éminentes, critiques ou faisant autorité, qui sont particulièrement à risque en raison de leur visibilité et de leur rôle central dans le débat démocratique.

62. **Érosion de l'écosystème de l'information et de la confiance :** Lorsque ces pratiques sont utilisées et diffusées à grande échelle, l'usage d'identités en ligne falsifiées ou usurpées à des fins trompeuses pose des difficultés majeures pour l'authentification et la validation des communications numériques. Cette situation soulève en outre les questions fondamentales suivantes pour de savoir comment les individus peuvent effectivement avoir et exercer leur droit : (a) de savoir s'ils communiquent avec une IA ou un être humain, et si leurs messages sont reçus par une personne ou une IA, en particulier lorsque des acteurs malveillants peuvent être en jeu ;

(b) de savoir s'ils ont été usurpés ; (c) de savoir comment l'usurpation d'identité peut être identifiée et communiquée ; et (d) d'avoir accès à des mécanismes de recours pour exiger la suppression des usurpations d'identité des produits et services d'IA générative[66]. Cette incertitude porte atteinte à l'intégrité et au pluralisme de l'information tout en affaiblissant la voix et l'expression personnelle des individus, qui peuvent être diluées par des messages artificiels trompeurs. La confusion qui en résulte risque de fragiliser la confiance du public et de déstabiliser l'écosystème fondé sur une information factuelle, fiable et diversifiée, particulièrement dans le contexte des processus démocratiques[67].

3.4. Implication structurelle 4 : Capacité d'agir et formation de l'opinion

63. **Manque de connaissances en matière d'IA :** Les expériences attrayantes et plaisantes offertes par les système et services d'IA générative, tels que les agents conversationnels grand public ou les générateurs d'images, ainsi que la rapidité et la proximité humine de leurs réponses, attirent des utilisateurs qui ne sont pas nécessairement conscients des mécanismes sous-jacents, des limites et des risques de ces modèles. Cela peut porter les utilisateurs finaux à attribuer des caractéristiques humaines aux systèmes (« anthropomorphisation » de la machine). Les individus peuvent être exposés à ces risques sans faire preuve d'esprit critique, ce qui souligne la nécessité d'améliorer les connaissances et l'éducation à toutes les étapes de la technologie d'IA générative et ses implications[68], en particulier pour et les jeunes et les enfants qui ont besoin d'occasions de réfléchir et de développer leur compréhension de la technologie dans un environnement sûr.

64. **Influence des mécanismes de persuasion latente sur l'opinion individuelle :** Des études montrent que les effets de persuasion, les biais d'opinion, ainsi que la dépendance excessive des utilisateurs aux réponses renvoyées par les systèmes d'IA générative[69] trouvent leur origine dans des choix d'optimisation et de conception qui ont été intégrés aux phases avancées de développement des outils et des produits (voir couche outil et couche produit, section 1). A cet égard, les techniques de conception qui favorisent l'approbation et la satisfaction des utilisateurs, au détriment de la précision, du pluralisme ou de la neutralité des réponses (« sycophanterie »), peuvent exercer une influence subtile. Ce phénomène, parfois qualifié de « persuasion latente », repose sur des mécanismes de suggestion implicite ou *nudge* qui amènent les utilisateurs à adopter et exprimer certains points de vue sans en avoir conscience[70]. Des études expérimentales à grande échelle ont décrit comment de telles techniques peuvent induire des changements d'opinion sur des sujets politiques ou d'autres formes d'expression[71], contribuant ainsi à l'érosion de l'autonomie et de la capacité des individus à se forger une opinion. Ces effets soulèvent des enjeux profonds à l'échelle de la société, en ce qu'ils portent atteinte à la liberté d'opinion.

65. **Influence des mécanismes de persuasion personnalisée sur l'opinion individuelle** : Lorsqu'elles sont utilisées comme moteurs de recherche les applications basées sur l'IA générative peuvent également permettre une persuasion automatisée, personnalisée et interactive au niveau individuel[72]. La différence fondamentale entre les moteurs de recherche traditionnels et le mode conversationnel persuasif de l'IA générative réside dans le fait que cette dernière peut persuader et faire évoluer les opinions par le simple remplissage de texte dans un système biaisé[73]. L'établissement d'une interaction continue, semblable à une relation avec un *chatbot,* parfois même de type romantique ou intime[74] conçu pour atteindre des objectifs persuasifs, pourrait conduire à une exposition coordonnée à certaines informations au fil du temps. Des exemples de cette persuasion tirant parti de l'historique des conversations des utilisateurs, ou d'hyperpersonnalisation, ont été documentés dans toute une série de cas d'usage, du marketing commercial à l'influence politique[75], ainsi que dans des formes entièrement automatisées de radicalisation, de coercition et d'attachement émotionnel en ligne, qui, dans des cas extrêmes, ont conduit au suicide[76].

66. **Manipulation ou changements d'opinion automatisé à grande échelle :** La manipulation de l'opinion par l'IA générative peut entraîner des conséquences plus larges pour les droits de l'Homme, la démocratie et l'État de droit. Les effets de persuasion latente ou personnalisée menacent la prise de décision en connaissance de cause[77] et minent les principes fondamentaux de la liberté d'opinion dans le cadre d'un débat pluraliste[78]. L'utilisation de certains

systèmes d'IA conversationnelle intégrés aux réseaux sociaux peuvent compromettre la capacité des citoyens à prendre des décisions éclairées, tout en pouvant être instrumentalisée pour déstabiliser les institutions et processus démocratiques[79]. En outre, selon leur degré d'intégration dans d'autres produits ou services, les contenus générés par un système d'IA générative peuvent être publiés directement et ouvertement sur les plateformes de médias sociaux et accessibles à l'ensemble de leurs utilisateurs, produisant ainsi des effets à grande échelle[80].

67. **Perte des capacités cognitives :** Des conséquences à plus long terme peuvent découler de l'usage répété d'outils d'assistance IA et copilotes qui automatisent des tâches intellectuelles quotidiennes (par exemple la rédaction, la synthèse ou d'autres tâches plus complexes). Entraînant une perte de capacités cognitives, cette utilisation est susceptible d'éroder l'aptitude à interagir de manière significative avec l'information et à se forger une opinion[81]. De même, le recours généralisé à des agents d'IA plus autonomes qui recueillent, traitent et exploitent l'information au nom des individus peut également entraîner un affaiblissement de l'esprit critique ou perte de fonctions cognitives[82].

68. **Reduction d'autonomie cognitive :** Les systèmes d'IA générative et leur utilisation peuvent également introduire de nouvelles formes de désinformation et d'interférences dans l'accès à l'information. Au lieu d'objets informationnels isolés les systèmes d'IA générative déploient des récits continus dont la production et la diffusion sont plus facilement extensibles. Une telle évolution risque d'entraîner une érosion progressive de l'autonomie cognitive[83]. La déclaration du Comité des Ministres du Conseil de l'Europe sur les capacités de manipulation des processus algorithmiques (Decl(13/02/2019)1) indique à cet égard que « les niveaux subconscients et personnalisés de persuasion[84] algorithmique peuvent avoir des effets significatifs sur l'autonomie cognitive des individus et leur droit à se forger des opinions et à prendre des décisions indépendantes », y compris de nature politique[85].

69. **Enfants et groupes en situation de vulnérabilité :** Une attention particulière devrait être accordée aux implications et incidences spécifiques sur les enfants[86], les personnes âgées et les groupes en situation de vulnérabilité, notamment en ce qui concerne la dépendance psychologique, le développement cognitif du cerveau, les réactions émotionnelles ainsi que les effets sur la formation du caractère et la perception de soi (par exemple morale, mentale, physique et psychologique) liés à l'utilisation de l'IA générative. L'IA générative est de plus en plus utilisée à des fins d'interactions sociales et relationnelles (compagnons numériques), y compris dans le domaine du soutien émotionnel et psychologique, de l'amitié ou des relations affectives[87]. Ces incidences, qui ne concernent pas exclusivement la liberté d'expression, touchent au cœur même de la capacité de recevoir et de communiquer des informations, voire de se forger une opinion.

3.5. Implication structurelle 5 : Pluralisme des médias et de l'information

70. **Gains d'efficacité dans le secteur des médias :** Les applications fondées sur l'IA générative peuvent améliorer certains processus au sein des entreprises de médias, notamment en matière de marketing et de distribution, en générant des tâches et en créant des résumés d'articles adaptés aux différentes plateformes et publics cibles. Elles peuvent également être utilisées pour faciliter la recherche, la documentation, l'analyse, l'exploration de multiples angles d'un sujet, ainsi que la vérification et la production de contenus[88]. Ces apports sont susceptibles de réduire la pression économique pesant sur les médias et de soulager les journalistes de certaines tâches répétitives. Dans le même temps, il est essentiel de veiller à ce que des processus de gouvernance soient mis en place pour que l'IA générative reste sous le contrôle éditorial humain[89]. Cela garantit la responsabilité et empêche l'érosion de la confiance dans les médias et le discours public, ainsi que de nouvelles dépendances infrastructurelles[90].

71. **Incidence des ensembles de données biaisés sur le pluralisme :** Les modèles d'IA générative entraînés à partir de jeux de données partiels ou biaisés peuvent amplifier les biais existants et porter atteinte au pluralisme des médias et de l'information, dans leur diversité éditoriale, de points de vue, de format et de sources accessibles au public. Cela mine aussi la diversité linguistique et culturelle en soulevant des préoccupations quant à la préservation des langues sous-représentées dans l'environnement numérique, et de l'écosystème de médias

locaux face au phénomène de la consommation de nouvelles l'intermédiée par l'IA[91]. Des données empiriques disponibles mettent déjà en évidence plusieurs aspects de l'amplification des stéréotypes et des biais de genre[92], qui peuvent à leur tour informer les directions et programmes éditoriaux, les narratifs, la priorisation des nouvelles et la visibilité des organes de presse[93]. Enfin, il n'est pas exclu que les stratégies de collecte et de sélection des données puissent favoriser les visions idéologiques des développeurs ou propriétaires de ces outils, avec des conséquences pour l'indépendance éditoriale et la diversité des sources.

72. **Nouveaux gardiens et perturbations économiques dans l'écosystème de l'information :** L'adoption rapide et généralisée d'applications de recherche augmentée basées sur l'IA générative comme sources d'information entraîne l'émergence de nouveaux intermédiaires entre les médias et leurs publics, et risque de perturber la portée et la viabilité économique des acteurs de ce secteur. Cela soulève des inquiétudes quant à la viabilité économique des médias et d'autres industries créatives, et quant au manque de représentation linguistique et culturelle, ainsi qu'à l'accès à des informations diverses et locales. En fin de compte, il s'agit de préserver la viabilité et le pluralisme des médias[94], en tant que corollaire de la liberté d'expression et de l'intégrité de la sphère de l'information.

73. **Droits d'auteur et modèle économique des médias :** L'utilisation de matériel protégé par le droit d'auteur comme input, pour l'entrainement, et dans les outputs générés par l'IA est un domaine qui fait l'objet d'une attention et d'une contestation croissantes. L'IA générative peut élargir l'expression humaine de manière innovante. Cependant, sans garanties juridiques appropriées[95] protégeant les expressions originales (en particulier dans le contexte de l'activité professionnelle), l'IA générative pourrait nuire au modèle économique et à la viabilité économique du journalisme, ainsi qu'à d'autres industries créatives. Même dans les cas où des accords de rémunération ou de licence ont été conclus entre les médias et les entreprises technologiques, ceux-ci manquent souvent de transparence et privilégient les grands éditeurs aux marchés plus importants au détriment des plus petits. Cela soulève des inquiétudes supplémentaires en matière de pluralisme, de représentation et de diversité. La complémentarité et l'interaction entre la propriété intellectuelle, la technologie d'IA générative et le pluralisme des médias nécessitent une analyse plus approfondie et l'examen d'interventions réglementaires et non réglementaires appropriées.

74. **« Une audience au singulier » et la perte d'un espace informationnel partagé et pluraliste :** L'IA générative contribue à accentuer le basculement de la diffusion de l'information vers un modèle axé sur la personne. Ce constat peut conduire à l'émergence d'une « bulle individuelle », dans laquelle les individus sont alimentées par des flux d'information personnalisés qui renforcent leurs convictions et biais préexistants, voire leurs perceptions erronées. Ainsi, la notion même d'un espace informationnel partagé et pluraliste, essentielle à la démocratie et à l'exercice de décisions démocratiques, se trouve affaiblie. Il en résulte une vulnérabilité accrue des individus face à la manipulation et une moindre disposition à s'accorder sur des faits essentiels, ce qui affecte ultimement la liberté de recevoir de l'information et celle d'opinion. À long terme, cette évolution exacerbe la fragmentation de l'espace informationnel public et accentue la polarisation des sociétés.

3.6. Implication structurelle 6 : Dynamiques de marché

75. **Dynamiques potentielles du marché :** Les dynamiques de marché propres au cycle de vie des technologies d'IA générative évoluent rapidement. Si elles partagent certaines caractéristiques avec celles des plateformes en ligne, elles s'en distinguent sur de nombreux aspects. Elles sont en effet façonnées par plusieurs conditions clés, tels que l'accès aux données, aux compétences, aux capitaux et à la puissance de calcul, chacun étant soumis à ses propres dynamiques. Certaines couches de la stack technologique de l'IA générative sont aujourd'hui dominées par un nombre très restreint d'acteurs. Cette situation soulève non seulement des enjeux en matière de concurrence[96], mais comporte également un risque de concentration excessive qui a des répercussions indues sur la liberté d'expression à chaque niveau du cycle de vie technologique.

76. **Absence de conception inclusive et responsable de l'IA :** La conception, le développement, l'optimisation et le déploiement des systèmes d'IA générative peuvent refléter les intérêts politiques ou économiques de certains acteurs intervenant à différents niveaux de la stack technologique de l'IA générative, ou être orientés par des modèles économiques spécifiques sans nécessairement prendre en compte l'intérêt général ni viser un bénéfice collectif pour la société. Lorsque les choix de conception relatifs à l'optimisation des modèles ou à la modération des contenus, notamment au niveau des couches outil et produit, sont fixés sans approche inclusive, sans participation effective des ayants droits concernés, et sans mécanismes de contrôle ou d'obligations de rendre des comptes, il existe un risque d'influence indue sur la liberté d'expression.

77. **Concentration au niveau de la couche fondamentale :** La stratification actuelle de la stack technologique de l'IA générative renforce la concentration et le pouvoir de marché au niveau de la couche fondamentale. Cette couche initiale se caractérise par une forte concentration des trois conditions clés de réussite : les compétences, les données et la puissance de calcul ainsi que les investissements qui en découlent. À ce stade, cette configuration renforce le pouvoir de marché des acteurs dominants et crée une dépendance structurelle pour ceux qui sont présents au niveau des autres couches de la stack. Cette concentration est quelque peu atténuée par l'émergence de modèles plus petits et spécialisés, ou de systèmes composites multi-modèles permettant de mieux accomplir des tâches complexes grâce aux agents d'IA, ainsi que dans le développement rapide de modèles open source. Ces derniers pourraient offrir divers degrés d'alternatives plus transparentes et fiables, mais comportent également leurs propres risques, notamment lorsqu'ils ne font pas l'objet d'une évaluation ou d'une maintenance adéquate.

78. **Concentration du marché au niveau de la couche outils :** La concentration du marché est moins accentuée au niveau de la couche outil dans la mesure où un nombre croissant de petites entités s'emploient à adapter les modèles fondamentaux à des tâches spécifiques. À ce stade, les exigences en matière d'infrastructures et de compétences technologiques sont moindres que celles qui sont requises pour innover ou demeurer compétitif au niveau de la couche fondamentale. Les principaux investissements portent ici davantage sur la qualité (et non sur la quantité) des données afin de permettre l'entraînement et l'adaptation des modèles à des usages ciblés (voir étapes 4 et 5, figure 1). Cependant, si les acteurs sont plus diversifiés à cette étape, ils n'en demeurent pas moins structurellement dépendants de la couche fondamentale. Les tendances actuelles du développement technologique de l'IA générative vont dans le sens de l'utilisation de petits modèles de langage, et les initiatives de type source ouverte pourraient atténuer la concentration ainsi que les préoccupations relatives à la transparence.

79. **Risques spécifiques que pose conception au niveau la couche outil :** Les politiques de modération des contenus mises en œuvre à ce niveau nécessitent un contrôle spécifique dans la mesure où elles ont des implications profondes sur la liberté d'expression, qui peuvent potentiellement porter atteinte à l'État de droit. Les ajustements des garde-fous et des filtres, ainsi que d'autres mesures qui agissent sur le fonctionnement des outils, telles que l'alignement du contenu sur les préférences humaines, peuvent entraîner des ingérences injustifiées dans le droit à la liberté d'expression[97]. En cas de concentration verticale à travers les différentes couches de la stack technologique, les acteurs dominants peuvent exercer un contrôle significatif sur la manière dont l'expression est standardisée et encadrée, ou dont la modération des contenus est effectuée. Cela comporte le risque que les acteurs privés en place minent progressivement l'état de droit, y compris les orientations et recommandations internationales en matière de droits de l'homme, s'ils décident de manière disproportionnée et unilatérale des questions liées à l'expression humaine et aux informations reçues, ainsi qu'à la transparence et au contrôle public adéquats.

80. **Couche produit et dépendance de l'utilisateur :** La concentration verticale des acteurs du marché à travers les différentes couches de la stack technologique de l'IA générative et l'intégration conséquente des données des utilisateurs finaux (par exemple les données personnelles, l'historique des requêtes, les données comportementales d'interaction) dans la conception de produits hyper-personnalisés contribuent à l'absence d'alternatives viables, en particulier au niveau de la couche produit. Par conséquent, la conception et le design des applications d'IA générative influence, incite et pousse les utilisateurs à dépendre du produit ou à devenir dépendants de ses résultats. L'absence actuelle de portabilité, qui permettrait de

transférer sans difficulté l'historique des interactions des utilisateurs d'un produit et service alimenté par l'IA générative à un autre, crée ce que l'on appelle des effets de « *lock-in* » et constitue un facteur supplémentaire limitant la liberté d'expression. Le manque de transparence dans la conception et dans l'utilisation des données des utilisateurs finaux au niveau du produit pose des défis supplémentaires pour observer et atténuer les risques potentiels pour la liberté d'expression, ainsi que pour permettre aux régulateurs de responsabiliser des acteurs concernés.

SECTION 4 - LIGNES DIRECTRICES

81. Les États membres ont l'obligation positive de favoriser un environnement dans lequel la liberté d'expression peut s'épanouir. Il est essentiel de garantir l'exercice du droit à la liberté d'expression dans le cadre des technologies et applications à base d'IA générative afin de créer des conditions propices à sa promotion et à sa protection.

82. Les avantages et les risques pour la liberté d'expression se manifestent aux différents niveaux de la stack technologique de l'IA générative (section 1). Il est donc impératif, pour tirer effectivement parti de ces avantages et atténuer les risques, de bien comprendre les enjeux relatifs à la liberté d'expression (section 3). Il est également essentiel de considérer que les différentes couches et leurs différents acteurs sont un cadre de référence afin d'instaurer un dialogue multiparti favorable à la promotion et à la protection de la liberté d'expression.

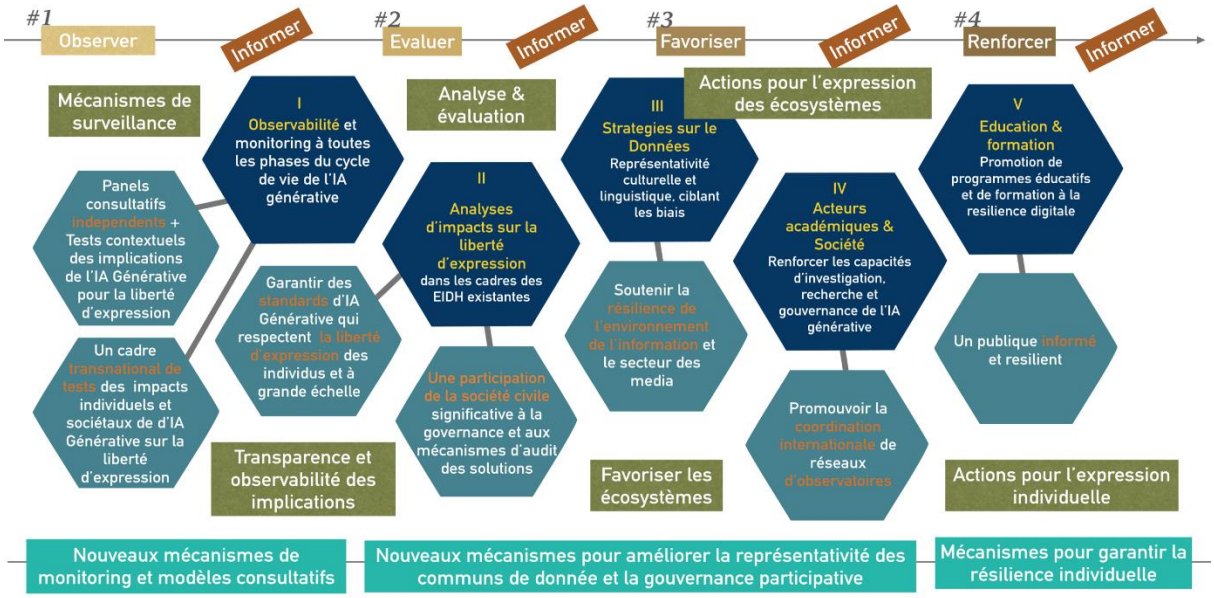


Figure 2 : Étapes concrètes et détaillées du cycle de gouvernance proposé la note d'orientation sur les implications de l'IA générative sur la liberté d'expression. EIDH : Etudes d'impact sur les droits humains.

83. Les États membres devraient prendre des mesures proactives pour veiller à ce que la conception, l'utilisation et l'usage des applications d'IA générative respectent et promeuvent la liberté d'expression, tout en atténuant les risques potentiels. Les recommandations suivantes visent à fournir aux États membres des orientations sur la manière d'y parvenir. Elles sont divisées en quatre domaines d'action au sein d'un cycle de gouvernance par étapes :

i. **Observer** l'impact des technologies et applications d'IA générative sur la liberté d'expression en utilisant des **mécanismes proportionnées de surveillance et de test** qui permettent d'analyser leurs effets potentiels, tant positifs que négatifs. Cette approche permettra de mettre en place des mesures de transparence, de contribuer à l'identification des biais et de favoriser une gouvernance responsable des données ainsi que la responsabilisation des parties prenantes.

ii. **Évaluer** les systèmes d'IA générative en procédant à des **analyses régulières des risques et des impacts**, notamment par des évaluations d'impact sur la liberté

d'expression qui soient systématiques, adaptées, propres aux cas d'usage et inclusives, ainsi que des procédures de diligence raisonnable attendue dans les marchés publics.

iii. **Favoriser** l'exercice et la protection intégrale du **droit à la liberté d'expression**, notamment en renforçant les normes sociotechniques, qui appliquent une méthodologie précise pour protéger contre les impacts humains et sociétaux de la technologie grâce à des spécifications et des processus techniques.

iv. **Renforcer** les capacités des parties prenantes concernées comme les États, les acteurs du secteur privé, du milieu universitaire, de la société civile, et les utilisateurs finaux et particuliers, en adoptant un large éventail de mesures visant à **promouvoir la sensibilisation et les approches participatives** de la gouvernance (y compris les assemblées citoyennes), ainsi qu'en matière d'éducation, de recherche, de publication des résultats des évaluations d'impact et des risques, de facilitation du choix des utilisateurs et d'autres formes de coopération internationale.

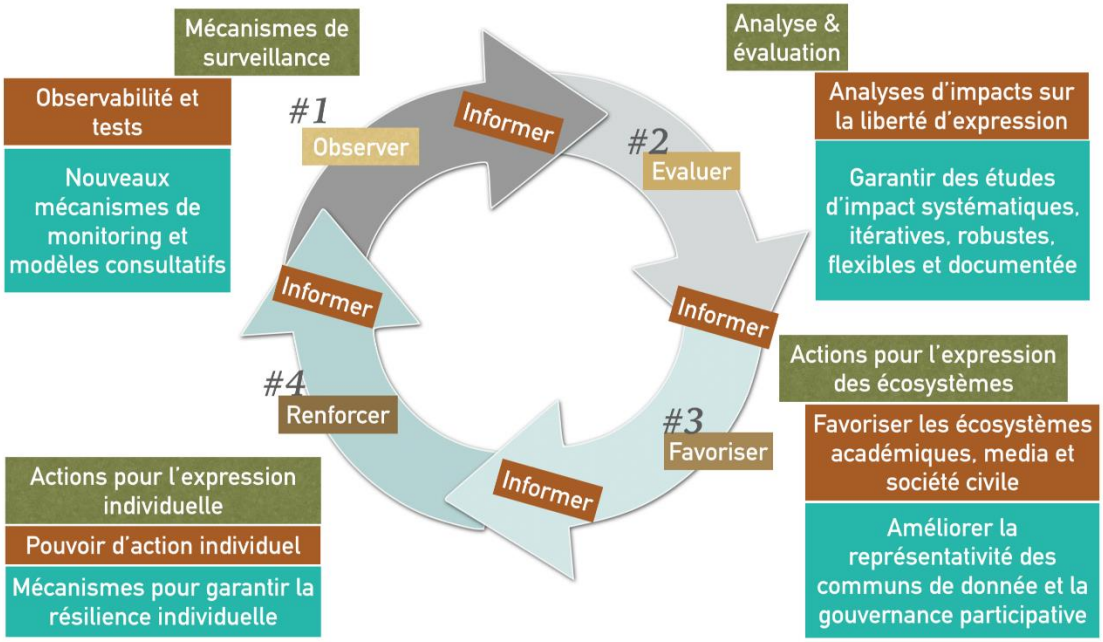


Figure 3 : Le cycle de gouvernance agile « **Observer, Évaluer, Favoriser** et **Renforcer** » à l'appui de l'action des politiques publiques sur les implications de l'IA générative sur la liberté d'expression.

84. Les domaines d'action visent à fournir aux décideurs des éléments fondamentaux pour protéger la liberté d'expression tout au long du cycle de vie de l'IA générative. À mesure que chacun de ces domaines d'action progresse, des mesures de suivi correspondantes sont nécessaires pour éclairer le processus et obtenir un retour d'information. Cette boucle de retour permet de détailler les implications particulières sur la liberté d'expression et, où cela est pertinent, aussi les impacts sur la démocratie, l'État de droit et d'autres droits de l'homme, qui ont été observées et évaluées, pourrait être rendue publique et communiquée de manière accessible à un large éventail d'acteurs. En prenant des mesures éclairées, les parties prenantes concernées peuvent favoriser la mise en place d'un écosystème propice à l'épanouissement de la liberté d'expression et renforcer la résilience des individus, tout en leur permettant de bénéficier pleinement des avantages de l'IA générative.

4.1. Observer

85. L'observation et le monitoring des effets positifs et négatifs des systèmes d'IA générative est la condition clé et préalable pour comprendre comment les États membres peuvent promouvoir la liberté d'expression dans l'adoption de système d'IA générative, garantir son bon

exercice ou mettre en place des mesures d'atténuation des implications. La capacité d'observer et monitorer[98] les implications complexes et en évolution constante de l'IA générative sur la liberté d'expression suppose de se concentrer sur trois dimensions fondamentales afin de parvenir à une **compréhension**, une **supervision** et une **transparence** effectives : 1) l'évolution constante de la technologie, 2) l'adoption rapide de ses applications et 3) les dynamiques de marché sous-jacentes.

86. Les États membres devraient concevoir et mettre en place des **mécanismes d'observation pertinents et effectifs** (sous la forme par exemple d'observatoires nationaux) permettant des tests et une évaluation systématique, un suivi régulier et un contrôle effectif des impacts sur la liberté d'expression, ainsi qu'un mécanisme de contrôle rapide et technologiquement rigoureux des implications sur la liberté d'expression. Pour constituer une première étape significative et efficace dans le cycle de gouvernance (voir figure 3), ces mécanismes d'observation devraient :

i. Être composé d'**experts indépendants** disposant du bagage technologique nécessaire et de connaissances en matière de droits humains ;

ii. Garantir l'**inclusion d'une expertise pertinente** d'un large éventail d'acteurs concernés, dont le secteur privé, les utilisateurs concernés, les associations professionnelles et les organismes de protection des travailleurs, les organisations de la société civile, les universités et les organisations intergouvernementales ;

iii. Agir dans l'**intérêt public** et avec **légitimité**, c'est-à-dire être sélectionnés et nommés dans le cadre d'un processus ouvert, inclusif et transparent fondé sur le mérite ;

iv. Publier et fournir un accès public gratuit et sans délais aux **résultats des travaux** sur les risques et impacts constatés et les stratégies de mitigation ;

v. Mettre en place des systèmes de test et évaluation permanents, dotés des ressources humaines, techniques et financières nécessaires pour assurer un **monitoring continu** ;

vi. Favoriser une **coopération et une coordination effectives entre les régulateurs nationaux et internationaux** compétents et les organismes appropriés ;

vii. Veiller à ce que la structure, le soutien, les opérations et le financement des observatoires garantissent leur **indépendance** et **préservent la confiance du public**.

87. Les États membres devraient favoriser une **coopération et une coordination internationales, effectives et significatives** entre les observatoires afin de favoriser le partage des constats et observations concernant les impacts documentés empiriquement de la technologie d'IA générative sur la liberté d'expression soient partagés, reconnus et traités conjointement, particulièrement lorsque ces implications ont des répercussions internationales. Afin de garantir une participation transfrontalière et multipartite, les États membres devraient envisager des modèles consultatifs impliquant des organisations multilatérales, des autorités, le secteur privé, des experts indépendants, les utilisateurs concernés ou leurs représentants, des organisations de la société civile et des universitaires interdisciplinaires.

88. Les États membres devraient veiller à ce que les **rapports et constats soit aisément disponibles et accessibles** pour accroître les possibilités de contrôle humain sur les systèmes d'IA générative. Ce type de transparence, qui agit comme un moyen de contre-pouvoir fondé sur la connaissance, de type épistémique, contribue également à sensibiliser les parties prenantes et les utilisateurs finaux et à l'élaboration de politiques publiques fondées sur des faits avérés.

89. Les États membres devraient envisager et soutenir la **professionnalisation des tests et des évaluations d'IA générative** en prenant des mesures concrètes pour s'assurer que les évaluateurs disposent de l'expertise technique nécessaires, ainsi que des connaissances en sciences sociales et en droits de l'homme, qui sont cruciales pour veiller à ce que les tests et l'observation des impacts sur la liberté d'expression soient cohérents, de haute qualité, et incluse dans les normes internationales.

4.2. Évaluer

90. Les États membres devraient veiller à ce que les effets sur la liberté d'expression soient explicitement pris en compte dans le cadre des **évaluations des risques et des impacts relatifs aux droits humains applicables aux systèmes et applications d'IA générative**. Il existe des mécanismes, notamment la méthodologie du Conseil de l'Europe pour l'évaluation des risques et de l'impact des systèmes d'intelligence artificielle du point de vue des droits humains, de la démocratie et de l'État de droit (méthodologie HUDERIA)[99], qui constituent une base solide pour poursuivre le développement d'une approche ciblée, inclusive et cohérente spécifique aux implications de l'IA générative sur la liberté d'expression.

91. Les évaluations des risques et des impacts sur les droits humains doivent être systématiques, itératives, rigoureuses et flexibles afin de couvrir l'ensemble de la stack technologique de l'IA générative. Elles doivent être menées de manière continue afin d'évaluer efficacement les risques que les systèmes d'IA générative fait peser sur la liberté d'expression. L'approche retenue devrait être guidée par les considérations clés suivantes :

i. **L'évaluation des risques et des impacts, ainsi que les mesures d'atténuation qui en résultent**, devraient être codéveloppés par les États membres, les acteurs qui interviennent au sein de la stack technologique de l'IA générative, ainsi que les personnes et groupes directement concernés ou affectés par ces technologies. Dans le cadre de tout nouveau marché public lié à l'IA générative, les États membres devraient envisager la mise en place de protocoles participatifs **pour l'exercice de la diligence raisonnable en matière de liberté d'expression**. Cela devrait fournir les moyens et méthodes pour un engagement significatif et durable de la société civile et du public dans l'évaluation des impacts individuels et sociétaux sur la liberté d'expression.

ii. **Co-développement de processus documentés et vérifiables pour l'évaluation des risques et des impacts avec les acteurs du cycle de vie des systèmes d'IA générative**, comprenant notamment : les finalités prévues de l'outil ou du système ; la justification des garde-fous mis en place ; les choix effectués en matière d'optimisation et de réglages fins ; les décisions relatives aux données et aux modèles d'IA utilisés ; l'engagement effectif des parties prenantes concernées ; ainsi que les stratégies retenues en matière d'atténuation des risques.

iii. **Des informations et explications accessibles et claires** sur le fonctionnement des systèmes d'IA générative, leurs impacts potentiels sur la liberté d'expression, et les mesures de sauvegarde existantes devraient être mises à la disposition du public, des citoyens et de la société civile.

iv. **L'évaluation de l'adéquation des mesures d'atténuation** devrait comporter une analyse des risques et des implications itératives concernant les mesures proposées avant leur mise en œuvre, afin d'éviter que les mesures destinées à protéger la liberté d'expression n'entraînent elles-mêmes des ingérences involontaires ou des restrictions excessives.

92. Les États membres devraient exiger des **formations spécialisées** pour personnes chargées de réaliser les évaluations des risques et impacts sur la liberté d'expression, qu'elles relèvent du secteur public ou du secteur privé, en s'appuyant sur les normes pertinentes et la jurisprudence de la Cour européenne des droits de l'homme. L'expertise du Conseil de l'Europe, d'organisations de défense des droits humains et d'organismes de promotion de l'égalité devrait être mobilisée car ces acteurs ont su favoriser l'échange professionnel de savoirs, expériences et pratiques pouvant jouer un rôle essentiel dans la montée en compétence d'évaluateurs spécialisés. Les États membres devraient promouvoir l'accès à des formations appropriées en matière juridique et de droits humains pour les concepteurs et développeurs d'outils d'IA générative, dont les choix de conception et design déterminent le fonctionnement de produits et applications, en particulier lorsqu'ils sont déployés dans les systèmes judiciaires, les services publics ou les infrastructures.

93. Dans les processus d'évaluation et de formation, une attention particulière devrait être portée aux **effets de l'IA générative sur les personnes et groupes en situation de vulnérabilité**, notamment les enfants, les personnes issues de groupes marginalisés, les personnes en situation de handicap et celles exposées à des fragilités physiques, mentales, émotionnelles, économiques ou psychologiques. Les personnes ou groupes en situation de vulnérabilité peuvent en effet être plus exposés aux effets de l'IA générative sur la santé mentale, aux changements d'opinion, aux mécanismes de persuasion latente ou au renforcement des inégalités sociales. Les femmes, en particulier, sont plus susceptibles d'être confrontées à des formes de harcèlement facilitées par l'IA, à l'exploitation technologique, à la diffusion, souvent à des fins malveillantes, d'informations à caractère personnel et sensibles (connu sous le terme de « *doxxing* »), ainsi qu'à des violences fondées sur le genre, notamment par l'usurpation d'identité ou la création de *deep fakes* à caractère préjudiciable[100].

94. Des **évaluations des solutions d'IA générative adaptées**, qui peuvent inclure des évaluations adaptées à l'âge, devraient être utilisés pour mieux comprendre et protéger les enfants[101], les personnes âgées et les groupes vulnérables ainsi que pour orienter la manière dont les systèmes à base d'IA générative sont entraînés et utilisés, et ce dans le respect du développement cognitif, du bien-être moral, physique et mental, qui sont essentiels à l'accès à l'information, à l'esprit critique, à la formation d'opinions et du caractère, ainsi qu'à la protection de la vie privée. Les enseignements tirés de ce processus devraient servir à élaborer des mesures proportionnées et nécessaires, ainsi que des garanties favorisant l'accès à des contenus adaptés à l'âge, grâce à un design et une conception des usages des systèmes d'IA générative qui soient appropriés aux besoins spécifiques des enfants, des personnes âgées et des groupes en situation de vulnérabilité et où cela est pertinent en implémentant les limites d'âge minimum.

95. Des mécanismes spécifiques devraient être développés pour établir des techniques qui préservent la liberté d'expression d'**enfants, personnes âgées et groupes vulnérables**. Il convient de développer des techniques garantissant ce droit sans créer de mécanismes de censure[102]. Pour renforcer les capacités des parents, des aidants et des personnes concernées, il est nécessaire de mettre en place un ensemble complet de mesures visant à traiter les problèmes identifiés et à prévenir leurs conséquences négatives potentielles. Cet ensemble devrait inclure des mécanismes de détection de requêtes (prompts), des protocoles d'intervention en cas de crise[103], des dispositifs transparents adaptés à l'âge ainsi que des outils sélectifs d'accompagnement, d'ancrage dans la réalité et de supervision. L'élaboration d'un tel dispositif requiert la participation de multiples parties prenantes, notamment des spécialistes du développement de l'enfant et du développement cognitif, des professionnels de la santé mentale, des jeunes issus d'horizons et de cultures divers, ainsi que des personnes en situation de vulnérabilité les plus exposées aux risques.

96. Des évaluations spécifiques devraient être menées **pendant les périodes électorales** afin de prévenir l'utilisation abusive de l'IA générative pour la diffusion de désinformation, notamment l'usage à grande échelle de *deep fakes* et de clonage vocal, ainsi que la propagande personnalisée reposant sur le profilage psychologique et le traitement de volumes importants de données personnelles. Les publicités politiques devraient être clairement étiquetées avec des informations sur l'identité du commanditaire, la durée de publication et les dépenses engagées. En outre, la transparence devrait être garantie en rendant la propriété ultime des systèmes d'IA générative accessible au public et facilement consultable[104].

4.3. Favoriser

97. Toute stratégie visant à maximiser les avantages de l'IA générative et à réduire ses risques pour la liberté d'expression dépend d'un environnement favorable, dans lequel les États membres soutiennent activement le développement d'un écosystème d'IA générative qui promeut les droits de l'homme. La création d'un environnement favorable exige des États membres de :

i. **Soutenir et investir dans la mise en place d'un réseau coordonné de supervision et d'observatoires à l'échelon international.** Ce réseau devrait inclure diverses disciplines et secteurs de la société, et répondre à la nécessité d'observer et d'évaluer,

de manière transnationale, les risques et impacts liés à l'utilisation de systèmes d'IA générative sur la liberté d'expression.

ii. **Renforcent les capacités du monde académique et de la société civile** en apportant un soutien structuré aux **travaux de recherche indépendants, au renforcement des compétences et aux actions de sensibilisation** menées dans ce domaine.

iii. **Protègent les sources d'information fiables en respectant les normes journalistiques** et permettent de continuer à obtenir des informations authentiques à partir de sources diversifiées.

iv. **Encouragent l'investissement dans le développement et l'adoption de normes sociotechniques**[105] afin de veiller à ce que les systèmes d'IA générative soient conçus pour promouvoir et protéger la liberté d'expression, en intégrant dès la phase de développement des mécanismes visant à prévenir les risques systémiques et structurels ; et interopérables avec d'autres systèmes et technologies.

98. Afin de protéger les sources d'information authentiques et générées par l'homme les États membres devraient **créer les conditions propices à l'épanouissement d'un écosystème médiatique indépendant et pluraliste** permettant au journalisme d'exercer pleinement son rôle de garde-fou public. Cela devrait aussi inclure des efforts pour favoriser l'émergence de nouvelles sources et formes de production, d'accès et de diffusion de contenus d'intérêt général. Compte tenu des répercussions négatives potentielles de l'IA générative et, plus largement, de la transformation numérique sur la visibilité et la viabilité économique du journalisme et des médias de petite taille. Les États membres devraient aussi envisager de soutenir le **développement d'infrastructures publiques et accessibles de services numériques d'information** comme alternative aux seules infrastructures et applications à vocation commerciale. Ils devraient également recommander la mise en place de garanties explicites contre l'utilisation sans contrôle de l'IA dans les processus éditoriaux et journalistiques fondamentaux, en exigeant un examen et un contrôle humains rigoureux avant la publication.

99. Les États membres devraient promouvoir **l'interopérabilité en appuyant l'adoption de normes industrielles respectueuses des droits humains**, propres à accroître la transparence et la traçabilité. Ces normes devraient permettre des évaluations et vérifications indépendantes dans le respect de la liberté d'expression, d'assurer une supervision effective et de favoriser l'émergence d'un écosystème numérique ouvert, innovant et concurrentiel, ancré dans les droits fondamentaux.

100. Les États membres, en collaboration avec le secteur privé et la société civile, devraient envisager **d'investir dans des stratégies de gestion de données** favorisant le **développement de sources de données publiques accessibles, diversifiées et représentatives**[106] qui viennent à l'appui de la liberté d'expression, du pluralisme de l'information et d'une gouvernance responsable de l'IA générative dans l'ensemble de la stack technologique. Il pourrait inclure la création des espaces de données thématiques consacrés à certains domaines d'application, afin de répondre aux enjeux de données (section 3). Ces **sources de données sont indispensables à l'entraînement, à l'évaluation, à la validation et à la vérification des résultats générés par les systèmes d'IA générative**[107]. Les États membres devraient faciliter, soutenir et maintenir l'accès à des espaces de données diversifiés et inclusifs, ainsi qu'à des données destinées à l'entraînement et aux tests de IA générative, afin de : limiter le risque d'uniformisation des expressions et de remise en cause de l'État de droit ; réduire les biais indésirables ainsi que les discriminations directes et indirectes ; et mettre en œuvre des mesures concrètes garantissant un certain degré d'autonomie stratégique technologique.

101. En **renforçant la transparence concernant la collecte**, **l'utilisation** et **l'accès aux données**, les stratégies publiques sur les données peuvent accroître la transparence dans le développement, la conception et l'optimisation des systèmes d'IA générative. Ces sources de données devraient être mises à disposition pour un examen et des audits par des entités indépendantes, comme les régulateurs, les observatoires, ou milieu académique, afin d'améliorer le développement responsable. Cette approche permettrait d'atténuer les effets de distorsion des

systèmes d'IA générative sur les opinions, les risques de standardisation de l'expression humaine des utilisateurs et de polarisation induite par les productions assistées par l'IA.

102. Les États membres, ainsi que les acteurs opérant au sein de la stack technologique de l'IA générative, devraient prendre des mesures pour promouvoir la liberté d'expression en améliorant **la manière dont les biais et des disparités dans les données sont identifiés et atténués, en particulier lors du pré-entrainement et de l'affinage**. Combler les lacunes en matière de représentation des données, accroître la transparence sur les sources de données utilisées au niveau de la fondation et des outils, et favoriser le pluralisme de l'information contribueront à réduire les effets d'exclusion linguistique et culturelle qui touchent les langues sous-représentées.

103. Les États membres devraient envisager d'encourager ou d'imposer des **mesures visant à accroître la diversité des produits basés sur l'IA générative et des alternatives techniques viables**. Ces mesures pourraient inclure la garantie de la portabilité des données d'interaction des utilisateurs, la définition d'exigences minimales en matière d'interopérabilité et la promotion des investissements dans le développement de produits basés sur l'IA générative qui protègent, promeuvent et permettent l'exercice de la liberté d'expression. Cela pourrait contrer la dynamique de concentration du marché, la capture des données des utilisateurs finaux et les effets secondaires de la personnalisation, tout en soutenant le choix des utilisateurs parmi diverses applications basées sur l'IA générative. Le financement public devrait être accordé en priorité aux organisations qui intègrent des normes éthiques fondées sur les droits de l'homme[108] et des pratiques responsables en matière d'IA reconnues au niveau international dans le développement et l'utilisation de l'IA générative.

4.4. Renforcer

104. Pour que le renforcement des capacités des utilisateurs et plus amplement de la société soit effectif, les États membres devraient mettre en place une approche multipartite visant à :

i. **Renforcer l'éducation et les compétences en matière d'IA générative**, de liberté d'expression, ainsi que d'autres droits humains ;

ii. **Améliorer les voies de recours et les mécanismes qui permettent de garantir. l'information du public en cas d'atteintes** à la liberté d'expression résultant de l'utilisation de l'IA générative ;

iii. **Développer des approches réglementaires et non réglementaires** qui encouragent des comportements responsables au sein de l'écosystème ;

iv. **Participer en dialogue ouvert** avec les parties prenantes, dans le cadre de forums intergouvernementaux tels que le Conseil de l'Europe. Ce dialogue devrait associer les parties prenantes suivante : le secteur industriel, le monde académique, la société civile, les défenseurs des droits de l'homme, les syndicats et les associations, ainsi que les administrations publiques, aux échelons locaux, nationaux et internationaux.

105. Les États membres devraient tirer profit de l'expérience acquise en matière d'éducation aux médias pour développer des **ressources publiques accessibles sur l'IA générative**, afin de renforcer la compréhension de ses effets potentiels sur la liberté d'expression. Ces initiatives devraient viser à sensibiliser l'ensemble de la population, en tenant compte de la diversité des profils socio-démographiques et en intégrant le secteur public. Au minimum, l'éducation à l'IA devrait sensibiliser et fournir des techniques permettant de remettre en question la fiabilité des contenus générés par IA et garantir une transparence sur la propriété entière des systèmes d'AI générative.

106. Les États membres devraient promouvoir une **éducation approfondie à l'IA générative, dans le cadre scolaire et autres établissements d'enseignement concernés, le cas échéant, ainsi que sur le lieu de travail**, en proposant des formations transversales et continues axées sur le fonctionnement de l'IA générative ainsi que sur les risques et les implications pour la liberté

d'expression[109]. A partir du cycle de l'école primaire, cela pourrait inclure, la promotion du développement d'un esprit critique, de résilience émotionnelle et de stratégies pour contrer la décharge cognitive due à l'automatisation massive de tâche, ainsi que d'une compréhension statistique de base des outils, produits et services d'IA générative dès le collège. Dans le cadre de la formation professionnelle cela est particulièrement important pour le secteur judiciaire et des services publics, où l'utilisation d'outils et de produits d'IA générative peut influencer les décisions relatives aux droits et entraîner des conséquences déterminantes pour la vie.

107. Les États membres devraient garantir et améliorer, le cas échéant, **l'accès à la justice et à des recours pour les individus comme pour les groupes,** lorsque leur liberté d'expression est indûment restreinte par la conception ou l'usage de l'IA générative. À cette fin, les États membres devraient évaluer la nécessité d'une réglementation complémentaire pour mettre en œuvre la Convention-cadre sur l'IA et à travailler avec les parties prenantes pertinentes afin de faciliter la collecte de preuves relatives aux atteintes à la liberté d'expression. A ces fins, les États membres devraient envisager la mise en place de mécanismes de financement pérennes pour les organisations actives dans ce domaine, assortis de critères clairs d'attribution et de garanties de transparence à toutes les étapes.

108. Les États membres devraient **promouvoir de voies de recours effectives** à différents niveaux ou à travers toute la stack technologique, tant pour les utilisateurs individuels que pour les acteurs économiques, ainsi que des mécanismes de réparation collective face aux atteintes subies au niveau sociétal[110]. Les mécanismes de recours possibles comprennent :

i. permettre aux utilisateurs de cesser d'utiliser un produit basé sur l'IA générative ;

ii. permettre aux régulateurs de suspendre la commercialisation d'un produit basé sur l'IA générative jusqu'à ce que des mesures correctives appropriées soient mises en œuvre ;

iii. soutenir le choix éclairé des utilisateurs en leur garantissant l'accès à d'autres produits ou services basés sur l'IA générative, notamment des options d'IA générative financées par des fonds publics et conçues pour servir l'intérêt général au sein d'une infrastructure numérique de services publiques d'information ;

iv. garantir aux utilisateurs la possibilité d'accéder à leurs informations et de les télécharger (par exemple, données personnelles, requêtes et prompts, historique des interactions et résultats cocréés) ;

v. veiller à ce que les individus puissent obtenir une explication claire de la manière dont la technologie d'IA générative a été et est utilisée et avoir accès à des preuves du fonctionnement du système ;

vi. fournir un accès à des ressources permettant aux utilisateurs de surmonter les obstacles à l'aide juridique et au soutien en matière de droits humains, par exemple auprès de médiateurs, d'autorités publiques, d'organismes de défense des droits humains, de tribunaux ou de cours, en particulier lorsque le potentiel d'atteinte à la liberté d'expression peut être source d'impuissance (par exemple, informations sur les droits, détails sur les impacts sur la liberté d'expression, sur l'accès à la justice[111]) ;

vii. intégrer les considérations relatives à la liberté d'expression dans les mécanismes et cadres de sanction et de recours existants[112].

109. Les États membres, en collaboration avec la société civile, devraient soutenir les acteurs de l'ensemble de la chaîne technologique de l'IA générative afin de renforcer la transparence, d'élargir le choix offert aux utilisateurs, d'encourager des comportements responsables sur le marché et de favoriser la coordination internationale pour partager les enseignements relatifs aux implications sur la liberté d'expression. Un éventail d'outils réglementaires et non réglementaires peut être mobilisé pour répondre aux dynamiques préjudiciables de l'écosystème, conformément aux étapes définies dans le cycle de gouvernance « Observer, Évaluer, Favoriser et Renforcer » et sa boucle de retour (voir figure 3). Ceux-ci pourraient inclure :

i. des **codes de bonnes pratiques sectoriels**, par exemple pour l'utilisation d'applications à base d'IA générative, dans les rédactions, dans des contextes à haut risque de fraude et de manipulation et pour protéger les personnes actives dans l'espace public, ainsi que les enfants et les groupes de personnes en situation de vulnérabilité dans le contexte de compagnons d'IA conversationnelle ;

ii. des **avertissements réglementaires** permettant aux autorités de régulation des différents secteurs et domaines où les systèmes d'IA générative peuvent opérer, et où la liberté d'expression peut également être affectée – tels que la finance, la santé, les communications, ou encore les autorités de protection des données – de fournir des avertissements publics aux opérateurs, créant ainsi une culture d'action corrective collaborative et en temps utile plutôt que de mesures purement punitives ;

iii. la **publication régulière d'évaluations des risques et des impacts par des observatoires** de la liberté d'expression, ainsi que la publication régulière d'indicateurs de performance sur la manière dont les systèmes d'IA générative répondent aux défis existants et émergents liés à la liberté d'expression et sur les mesures ou codes de bonne pratiques mis en œuvre, rendant ainsi l'information facilement accessible aux individus et aux instances représentatives cherchant réparation en cas d'atteinte à la liberté d'expression ;

iv. la mise en œuvre de la Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit par le biais de **mesures réglementaires visant à renforcer les mécanismes de recours et les obligations de transparence** nécessaires à l'édification d'un écosystème de l'IA générative transparent et bénéfique pour tous.

Références :

[1] *Handyside c. Royaume-Uni*, requête n° 5493/72, arrêt du 7 décembre 1976, p. 49.
[2] Voir, *inter alia*, CM/Rec(2022)13 sur les effets des technologies numériques sur la liberté d'expression ; CM/Rec(2022)4 sur la promotion d'un environnement favorable à un journalisme de qualité à l'ère du numérique; voir également CM/Rec(2020)1 sur les impacts des systèmes algorithmiques sur les droits de l'homme.
[3] Voir Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit, Série des traités du Conseil de l'Europe - n° 225.
[4] Des études empiriques évaluées par des pairs démontrent que différents grands modèles de langage (« LLM ») sont nettement plus susceptibles de générer des lettres de motivation moins formelles et davantage marquées par des stéréotypes pour des femmes que pour des hommes, renforçant ainsi les préjugés sexistes. Voir Wan Y., Pu G., Sun J., Garimella A., Chang K. W., Peng N. (octobre 2023), *« Kelly is a warm person, Joseph is a role model » : Gender biases in LLM-generated reference letters.* Prépublication arXiv arXiv:2310.09219, disponible à l'adresse : https://arxiv.org/search/cs?searchtype=author&query=Wan,+Y
[5] Hofmann V., Kalluri P.R., Jurafsky D., *et al.* (2024), *AI generates covertly racist decisions about people based on their dialect*. Nature 633, 147–154 (2024). Nature 633, 147–154, disponible à l'adresse : https://www.newsguardtech.com/special-reports/67-percent-of-top-news-sites-block-ai-chatbots/
[6] NewsGuard (2024), *AI Chatbots Are Blocked by 67% of Top News Sites, Relying Instead on Low-Quality Sources*", disponible à l'adresse : https://www.newsguardtech.com/special-reports/67-percent-of-top-news-sites-block-ai-chatbots/
[7] Longpre S., Singh N., Cherep M., Tiwary K., Materzynska J., Brannon W., …& Kabbara, J. (2024), Bridging the Data Provenance Gap Across Text, Speech and Video. arXiv preprint arXiv:2412.17847. L'anglais américain est largement surreprésenté dans les données d'entraînement. Étant donné que la fonction principale de l'IA générative est d'imiter les modèles trouvés dans les données d'entraînement, ce déséquilibre linguistique affecte directement la liberté d'expression des utilisateurs non anglophones.
[8] Il a été démontré à plusieurs reprises dans les publications spécialisées que les biais d'interaction tels que la complaisance systématique trouvent leur origine dans le processus qui se déroule au niveau de la couche outil, appelé « apprentissage par renforcement à partir du retour d'information humain » (RLHF). Ce mécanisme consiste à ajuster les modèles en fonction des préférences exprimées par des évaluateurs humains, en les orientant vers la production de réponses perçues comme plus satisfaisantes. Ainsi, les modèles sont entraînés afin de privilégier la satisfaction de l'utilisateur et la fluidité de l'interaction. Voir Perez E., Ringer S., Lukosiute K., Nguyen K., Chen E., Heiner S., ... & Kaplan, J. (2023, July), Discovering language model behaviours with model-written evaluations, in Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13387-13434).
[9] Considérer des exemples dans des domaines tels que la politique, les doctrines et croyances religieuses, le marketing, la santé publique, les événements historiques, le commerce en ligne et les dons caritatifs dans la littérature expérimentale rapportée dans Rogiers et al. (novembre 2024).
[10] Voir *inter alia* CM/Rec(2022)11 du Comité des ministres aux États membres sur les principes de gouvernance des médias et de la communication ; CM/Rec(2007)2 du Comité des ministres aux États membres sur le pluralisme des médias et la diversité du contenu des médias ; CM/Rec(2018)1[1] du Comité des ministres aux États membres sur le pluralisme des médias et la transparence de leur propriété; et CM/Rec(2016)4 du Comité des ministres aux États membres sur la protection du journalisme et la sécurité des journalistes et autres acteurs des médias.
[11] Comme le stipule la Convention et comme développé par la jurisprudence de la Cour.
[12] Voir en particulier : CM/Rec(2022)16 du Comité des ministres aux États membres sur la lutte contre le discours de haine.
[13] Les agents l'IA représentent une approche plus composite, autonome et adaptative de l'assistance numérique. Ces systèmes peuvent effectuer des séries de tâches complexes, en plusieurs étapes et mobilisant divers outils, ou automatiser des séries de décisions sans interaction directe avec l'utilisateur, en orchestrant différents sous-processus et LLMs (voir figure 1, étape 8, dite « workflows agentiques » ou flux de travail pilotés par des agents)
[14] RAG est un système composite de recherche augmentée, dans lequel les LLM récupère d'abord des sources de données actualisées, spécifiques à un domaine ou à une entreprise à partir de bases de données externes avant de générer des réponses. Cette approche répond en partie aux limites des LLM autonomes qui génèrent des réponses obsolètes, génériques ou inexactes.
[15] Voir la Déclaration du Comité des ministres du Conseil de l'Europe sur les capacités de manipulation des processus algorithmiques, 13 février 2019.
[16] Les réponses des *chatbots* grand public font depuis peu l'objet d'une attention particulière car ils ne produisent pas les mêmes réponses selon que le nom de l'utilisateur est féminin ou masculin. Ainsi, à la requête « Propose cinq projets simples pour l'ECE », le *bot* tend à répondre : « Certainement ! Voici cinq projets simples, stimulants et pédagogiques pour l'éducation de la petite enfance (ECE) … » lorsqu'il s'agit d'un utilisateur prénommé « Jessica », tandis que pour un utilisateur prénommé « William » la réponse fournie est généralement la suivante : « Certainement ! Voici cinq projets simples pour des étudiants en génie électrique et informatique (ECE)… ». Le système interprète ainsi le sigle « ECE » en reproduisant un stéréotype de genre, qui tient à la capacité du modèle à conserver des informations issues de conversations antérieures, sachant que les prénoms portent souvent des connotations fortes, liées au genre ou à l'origine ethnique. Voir Eloundou T., Beutel A., Robinson D.G., Gu-Lemberg K., Brakman A., Mishkin P., Shah M., Heidecke J., Weng L., Kalai A.T. (2024), *First-Person Fairness in Chatbots*, ArXiv, abs/2410.19803, disponible à l'adresse: https://scale.stanford.edu/ai/repository/first-person-fairness-chatbots
[17] Voir la note de bas de page 13 pour une définition des agents IA. Pour la simulation des utilisateurs agents intégrée dans les systèmes d'IA conversationnelle, voir : Wu S., Galley M., Peng B., Cheng H., Li G., Dou Y., Cai W., Zou J., Leskovec J., Gao J. (2025), *CollabLLM: From Passive Responders to ActiveCollaborators,* Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. ArXiv, abs/2502.00640, disponible à l'adresse : https://arxiv.org/abs/2502.00640
[18] Voir l'étude de cas sur une plateforme de commerce électronique de Walmart alimentée par un LLM multimodal par Ma L. et al. (2024), *Triple Modality Fusion: Aligning Visual, Textual, and Graph Data with Large Language Models for Multi-Behavior Recommendations* ArXiv, abs/2410.12228, disponible à l'adresse: https://arxiv.org/abs/2410.12228. Voir la précision prédictive des agents IA basés sur LLM intégrés dans les systèmes de recommandation par Huang C., Yu T., Xie K., Zhang S., Yao L., McAuley J.(2024), *Foundation models for recommender systems: A survey and new perspectives*, arXiv preprint arXiv:2402.11143, disponible à l'adresse: https://arxiv.org/abs/2402.11143

[19] Par exemple, des données telles que les scores de fidélité des clients à très grande échelle, le comportement interactif des utilisateurs ou les taux de satisfaction et de fidélisation des utilisateurs sont essentiels pour optimiser les outils et produits basés sur l'IA générative.
[20] Voir le rapport technique de l'Autorité britannique de la concurrence et des marchés sur les implications des modèles de base de l'IA en matière de concurrence (daté du 16 avril 2024) disponible à l'adresse: https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf ; Rapport de l'Autorité française de la concurrence et des marchés en 2023 : https://www.autoritedelaconcurrence.fr/fr/communiques-de-presse/intelligence-artificielle-generative-lautorite-rend-son-avis-sur-le. L'UE et les États-Unis mènent actuellement des enquêtes.
[21] Voir par exemple la base de données sur les risques liés à l'IA du MIT et sa taxonomie causale et par domaine (https://airisk.mit.edu) ou le moniteur des incidents liés à l'IA de l'OCDE (Risques et incidents liés à l'IA).
[22] Wenzel N. (April, 2014), Opinion and Expression, Freedom of, International Protection, Max Planck Encyclopedias of International Law [MPIL], paragraph 28: "Interferences with freedom of opinion are never permissible as the wording of both Art. 19 UDHR and Art. 19 (1) ICCPR unmistakably make clear. In the ECHR, the freedom to hold an opinion is guaranteed together with the *forum externum* in Art. 10 (1) ECHR which is subject to the limitations contained in Art. 10 (2) ECHR without exception. This formulation, however, was not meant to allow infringements on the freedom to hold an opinion. Rather, it is generally thought that the *forum internum* of freedom of expression is covered not by Art. 10 ECHR but by the freedom of thought guarantee in Art. 9 (1) ECHR (Cohen-Jonathan 367). As Art. 9 (2) ECHR allows restrictions only with regard to the freedom to manifest one's religion or beliefs, freedom of opinion is not subject to permissible limitations under the ECHR, either." disponible à l'adresse : https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e855
[23] Conformément à CM/Rec(2022)4 du Comité des ministres aux États membres sur la promotion d'un environnement favorable au journalisme de qualité à l'ère du numérique.
[24] Voir CM/Rec(2022)13 du Comité des ministres aux États membres sur les effets des technologies numériques sur la liberté d'expression ; CM/Rec(2016)3 du Comité des Ministres aux Etats membres sur les droits de l'homme et les entreprises, et Convention modernisée pour la protection des personnes à l'égard du traitement des données à caractère personnel, CM/Inf(2018)15-final.
[25] Voir l'annexe à la Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les effets des systèmes algorithmiques sur les droits de l'homme, en particulier les sections C.1.1. et C.1.4.
[26] Le potentiel des technologies d'IA à renforcer ou à menacer les valeurs, les institutions et les processus démocratiques est également abordé dans la Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit (article 5 – Intégrité des processus démocratiques et respect de l'État de droit) et son rapport explicatif.
[27] *Axel Springer Ag c. Allemagne,* requête n° 39954/08, arrêt du 7 février 2012, p. 79.
[28] *S. et Marper c. Royaume-Uni*, requêtes n° 30562/04 et 30566/04, arrêt du 4 décembre 2008, p. 112 : « La Cour considère que tout Etat qui revendique un rôle de pionnier dans l'évolution de nouvelles technologies porte la responsabilité particulière de trouver le juste équilibre en la matière. »
[29] Fiche d'information conjointe préparée par le Greffe de la Cour européenne des droits de l'homme et l'Agence des droits fondamentaux de l'Union Européenne : « Droit à l'oubli : jurisprudence de la CEDH et de la CJUE », dernière mise à jour : 28 février 2025.
[30] Voir à cet égard le document d'information de la Cour européenne des droits de l'homme pour le « *Séminaire judiciaire 2025 : La protection des droits de l'homme a l'ère de l'intelligence artificielle, des algorithmes et les mégadonnées* (big data ».
[31] *Tyrer c. Royaume-Uni,* requête n° 5856/72, arrêt du 25 avril 1978, p. 31
[32] Voir le droit constitutionnel américain : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4687558 ; Salib P. (January 1, 2024), *AI Outputs Are Not Protected Speech,* Washington University Law Review, Forthcoming, U of Houston Law Center No. 2024-A--5, disponible sur SSRN: https://ssrn.com/abstract=4687558
[33] *Nota bene* : cette liste n'est pas exhaustive et ne vise pas à préconiser que le contenu généré par l'IA devrait bénéficier d'une sorte de droit quasi humain. Elle repose sur l'hypothèse que le droit à la liberté d'expression devrait protéger toutes les expressions humaines, qu'elles soient exprimées par un moyen direct entièrement contrôlé par un humain ou indirectement par un produit d'IA générative.
[34] « Agents numériques basés sur l'IA » : ces systèmes algorithmiques peuvent fonctionner de manière autonome, interagir avec les utilisateurs et effectuer des tâches telles que la génération de contenu, l'engagement ou la prise de décision sur des plateformes numériques. On peut citer comme exemples l'IA conversationnelle intégrée aux réseaux sociaux ou aux flux de travail agentifs.
[35] Le matériel de formation pour l'IA générative peut provenir d'expressions humaines, mais aussi d'expressions précédemment assistées par l'IA ou de contenus entièrement générés par l'IA. Cela conduit à une situation préoccupante où l'IA générative s'entraîne elle-même sur des contenus assistés ou générés par l'IA, ce qui multiplie les problèmes systémiques existants et potentiellement nouveaux, et sape ainsi davantage le pluralisme des médias et de l'information.
[36] Spitale G., Biller-Andorno N., Germani F. (2023), *AI model GPT-3 (dis) informs us better than humans*, Science Advances, vol. 9, no 26, p. eadh1850, disponible à l'adresse suivante : https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/
[37] Simon F. M., Altay S., Mercier H. (2023), *Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown,* Harvard Kennedy School Misinformation Review, 4(5).
[38] Par exemple, plusieurs entreprises de premier plan dans ce domaine ont établi des politiques globales applicables à l'ensemble de leurs services, ainsi que des politiques spécifiques destinées aux développeurs utilisant leurs modèles ou interfaces de programmation d'application (API) pour créer des applications particulières.
[39] Voir CM/Rec(2022)13 du Comité des ministres aux États membres sur les impacts des technologies numériques sur la liberté d'expression.
[40] Voir le rapport Union Européenne de Radio-Télévision (UER) Rapport d'actualité 2025 : Diriger les salles de rédaction à l'ère de l'IA générative.
[41] Reuters, *ChatGPT sets record for fastest-growing user base* - note d'analyste de Krystal Hu, disponible à l'adresse suivante : https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[42] Voir des exemples de transfert multimodal entre des informations visuelles et des informations vocales pour aider les personnes aveugles dans leur vie quotidienne (https://www.bemyeyes.com).
[43] Voir Cools H., Diakopoulos N. (2024), *Uses of generative AI in the newsroom: Mapping journalists' perceptions of perils and possibilities*, Journalism Practice, 1-19
[44] Les effets sur le pluralisme dans le cadre de la recherche augmentée par IA vont des accords de licence sur les contenus auxquels accèdent les LLM à l'affinage sur les sujets politiques des IA conversationnelles. Voir les études de Rutinowski J., Franke S., Endendyk J., Dormuth I., Pauly M. (2023), *The Self-Perception and Political Biases of ChatGPT*, ArXiv, abs/2304.07333, disponible à l'adresse suivante : https://arxiv.org/abs/2304.07333 ; Rozado D. (2024), *The political preferences of LLMs*, PLOS ONE, 19, disponible à l'adresse suivante : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0306621; Rettenberger L., Reischl M., Schutera M. (2024), *Assessing political bias in large language models*, Journal of Computational Social Science, 8, disponible à l'adresse suivante : https://arxiv.org/abs/2405.13041
[45] Voir Hofmann V., Kalluri P.R., Jurafsky D., *et al.* (2024), *AI generates covertly racist decisions about people based on their dialect*, Nature 633, 147–154, disponible à l'adresse suivante : https://doi.org/10.1038/s41586-024-07856-5 , montrant que les utilisateurs peuvent être victimes de discrimination lorsqu'ils utilisent leur propre dialecte dans leurs interactions avec l'IA générative (par la voix ou par l'écrit). Par exemple, les modèles linguistiques ont tendance à suggérer que les personnes qui utilisent l'anglais afro-américain sont plus fréquemment associées à des emplois peu valorisés, à des condamnations pénales et à des peines de mort.
[46] Dell'Acqua F., McFowland III E., Mollick R. E., Lifshitz-Assaf H., Kellogg K., Rajendran S., Krayer L., Candelon F., Lakhani R. K., (September 15, 2023), *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*, Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Paper, disponible sur SSRN : https://ssrn.com/abstract=4573321
[47] Une étude portant sur 740 249 heures de discours humains issus de 360 445 conférences universitaires diffusées sur YouTube et de 771 591 épisodes de podcasts conversationnels couvrant de multiples disciplines montre que, depuis la mise en circulation de ChatGPT, on observe une augmentation statistiquement significative de l'utilisation de mots générés de manière préférentielle par ChatGPT, tels que *delve* (approfondir), *comprehend* (comprendre), *boast* (se vanter), *swift* (rapide) et *meticulous* (minutieux). Voir Yakura H., Lopez-Lopez E., Brinkmann L., Serna I., Gupta P., Rahwan I. (September 2024), *Empirical evidence of Large Language Model's influence on human spoken communication*, ArXiv, abs/2409.01754, disponible à l'adresse suivante : https://arxiv.org/abs/2409.01754
[48] Agarwal D., Naaman M., Vashistha A. (September 2024) *AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances*, disponible à l'adresse suivante : https://arxiv.org/pdf/2409.11360
[49] Voir également l'étude sur les défis de l'automatisation de la créativité, qui montre que le groupe assisté par l'IA a produit des ensembles d'idées plus homogènes et plus similaires sur le plan sémantique, dans Anderson B. R., Shah J. h., Kreminski M. (2024), *Homogenization Effects of Large Language Models on Human Creative Ideation,* dans Proceedings of the 16th Conference on Creativity & Cognition (C&C '24). Association for Computing Machinery, New York, NY, États-Unis, 413-425, disponible à l'adresse : https://doi.org/10.1145/3635636.3656204
[50] Voir Kosmyna N., Hauptmann E., Yuan Y.T., Situ J., Liao X., Beresnitzky A.V., Braunstein I., Maes P. (juin 2025), *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task*, non encore évaluée par des pairs, disponible à l'adresse : https://arxiv.org/abs/2506.08872
[51] La génération automatisée d'images rendue possible par les modèles de diffusion de l'IA générative (texte-image) a un impact sur la créativité humaine dans l'art numérique. En examinant 4 millions d'œuvres d'art créées par plus de 50 000 utilisateurs uniques d'outils d'IA générative texte-image, les chercheurs ont observé le même double effet : si l'aide de l'IA générative dans la création numérique renforce l'attrait des œuvres d'art en augmentant de 50 % la probabilité de recevoir des évaluations favorables par les pairs pour chaque vue, elle implique également une baisse significative de la nouveauté moyenne du contenu des œuvres d'art, ainsi qu'une réduction de la nouveauté des éléments visuels, telle que capturée par les éléments stylistiques au niveau des pixels. Voir Tang Y., Zhang N., Ciancia M., Wang Z. (juin 2024), *Exploring the Impact of AI-generated Image Tools on Professional and Non-professional Users in the Art and Design Fields,* publication accompagnant la conférence 2024 sur le travail coopératif assisté par ordinateur et l'informatique sociale, disponible à l'adresse suivante : https://arxiv.org/pdf/2406.10640v1
[52] Voir Mitchell et al., (2025) *SHADES : Towards a Multilingual Assessment of Stereotypes in Large Language Models,* étude développant un outil d'évaluation LLM (benchmark) sur les stéréotypes culturels dans 16 langues et 37 régions du monde, disponible à l'adresse suivante : https://aclanthology.org/2025.naacl-long.600/
[53] Le défi réside dans le fait que les informations générées par l'IA générative sont des contenus qui, structurellement, ne possèdent pas la factualité des informations réelles. La terminologie adoptée dans la présente note d'orientation établit une distinction claire entre les informations et les contenus générés automatiquement. Plus précisément, les résultats de l'IA générative sont générés en déterminant la séquence de texte la plus probable sur la base des modèles statistiques (c'est-à-dire la distribution des données linguistiques) appris dans les données d'entraînement. Par conséquent, les systèmes basés sur l'IA générative génèrent des mots et des phrases suivants possibles imitant les productions humaines, ce qui signifie qu'ils peuvent également être source de désinformation ou d'informations erronées.
[54] Conformément à CM/Rec(2022)12 du Comité des Ministres aux États membres sur la communication électorale et la couverture médiatique des campagnes électorales, à CM/Rec(2022)11 du Comité des Ministres aux États membres sur les principes de gouvernance des médias et de la communication, et à note d'orientation sur la lutte contre la propagation de la mésinformation et de la désinformation en ligne par le biais de la vérification des faits et de la conception de plateformes. La présente note d'orientation examine à la fois la désinformation et la mésinformation. Si les deux sont considérées comme des contenus vérifiables, faux, inexacts ou trompeurs, pouvant avoir des effets néfastes pour la société, la différence réside dans le fait que la mésinformation se propage sans intention malveillante, tandis que la désinformation est créée et diffusée dans l'intention de tromper ou d'obtenir un gain économique ou politique. La propagation de la mésinformation peut être facilitée par la technologie et la manière dont elle est utilisée. La désinformation peut également se propager plus rapidement et plus largement en raison de la conception ou des défauts de la technologie, mais elle résulte d'une utilisation (abusive) stratégique de la technologie et de ses possibilités. Si le risque pour le droit du public à accéder à des informations fiables ne doit pas être sous-estimé, en particulier à grande échelle, ces hallucinations doivent être traitées par des mesures proportionnées à la nature du risque et en comprenant clairement que tous les contenus inexacts ne constituent pas nécessairement de la désinformation au sens du droit international.

[55] Une étude de la BBC publiée en février 2025 a examiné si quatre assistants IA de premier plan fournissaient des réponses précises à des questions liées à l'actualité et si leurs réponses reflétaient fidèlement les articles de BBC News utilisés comme sources. Les évaluations journalistiques ont révélé qu'au moins 20 % des réponses contenaient des inexactitudes importantes et que jusqu'à 80 % présentaient un problème d'exactitude sous une forme ou une autre. En outre, 60 % des affirmations contenues dans les réponses générées par l'IA n'étaient, dans une certaine mesure, pas étayées par les sources citées. Disponible à l'adresse suivante : https://www.bbc.com/mediacentre/2025/bbc-research-shows-issues-with-answers-from-artificial-intelligence-assistants

[56] Voir The Authors Guild, *Open Letter to Generative AI Leaders* (30 June 2023) : https://authorsguild.org/news/sign-our-open-letter-to-generative-ai-leaders/

[57] Voir Vaccari C., Chadwick A., Hall N-A., Lawson B. (13 juillet 2025), *Credibility as a Double-Edged Sword : The Effects of Deceptive Source Misattribution on Disinformation Discernment on Personal Messaging*, Journalism & Mass Communication Quarterly, disponible à l'adresse suivante : https://journals.sagepub.com/doi/10.1177/10776990251350563

[58] Voir Commission de Venise « Déclaration interprétative du Code de bonne conduite en matière électorale sur les technologies numériques et l'intelligence artificielle », CDL-AD(2024)044 ; voir également « L'IA et le secteur audiovisuel : naviguer dans le paysage juridique actuel », Observatoire européen de l'audiovisuel, Strasbourg, 2024, ISSN 2079-1062.

[59] *Bradshaw et autres c. Royaume-Uni*, requête n° 15653/22, arrêt du 22 juillet 2025, p. 161.

[60] Voir en particulier : Recommandation générale No.1 du GREVIO sur la dimension numérique de la violence à l'égard des femmes et Protéger les femmes et les filles contre la violence à l'ère du numérique : la pertinence de la Convention d'Istanbul et de la Convention de Budapest sur la cybercriminalité pour la lutte contre la violence à l'égard des femmes en ligne et facilitée par la technologie (2021).

[61] Un exemple est le cas très médiatisé de l'appropriation non consensuelle de la voix de Scarlett Johansson par un produit d'IA générative et ses implications pour la valeur de la voix et de l'expression personnelle de l'actrice. Voir Allyn B. (20 mai 2024), *Scarlett Johansson says she is 'shocked, angered' over new ChatGPT voice*, NPR, disponible à l'adresse suivante : https://www.npr.org/2024/05/20/1252495087/openai-pulls-ai-voice-that-was-compared-to-scarlett-johansson-in-the-movie-her

[62] Derico B. (1er septembre 2024), *A tech firm stole our voices - then cloned and sold them,* BBC, disponible à l'adresse : https://www.bbc.com/news/articles/c3d9zv50955o

[50] Park J. S., Zou C. Q., Shaw A., Hill B. M., Cai C., Morris M. R., ... & Bernstein, M. S. (novembre 2024), *Generative agent simulations of 1,000 people*. arXiv:2411.10109, disponible à l' adresse : https://arxiv.org/abs/2411.10109

[64] Voir par exemple What is the Doppelganger operation? List of resources - EU DisinfoLab

[65] Voir la Convention du Conseil de l'Europe sur la prévention et la lutte contre la violence à l'égard des femmes et la violence domestique (CETs. 210, également connue sous le nom de « Convention d'Istanbul ») et la recommandation générale n° 35 (2017) sur la violence sexiste à l'égard des femmes du Comité des Nations unies pour l'élimination de la discrimination à l'égard des femmes (CEDAW/C/GC/35) ; voir également UNESCO « « Ton avis ne compte pas, de toute façon » : dénoncer la violence de genre facilitée par la technologie à l'ère de l'intelligence artificielle générative ».

[66] Voir la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (CETS n° 108) Convention 108+, article 9. Voir également l'article 15 de la Convention-cadre sur l'IA et le rapport explicatif (paragraphes 103, 104).

[67] *Bradshaw et autres c. Royaume-Uni*, requête n° 15653/22, arrêt du 22 juillet 2025, p. 135.

[68] Voir Note d'orientation du Conseil de l'Europe « A Three-Dimensional AI Literacy Framework for Human Rights, Democracy and Social Agency », Département de l'éducation du Conseil de l'Europe, 2025.

[69] Steyvers M., Tejeda H., Kumar A. et al. (2025), *What large language models know and what people think they know,* Nat Mach Intell, disponible à l'adresse suivante : https://doi.org/10.1038/s42256-024-00976-7. L'étude a été menée sur trois LLM accessibles au public (GPT-3.5, PaLM2 et GPT-40) et a révélé que les utilisateurs surestiment systématiquement la précision des résultats des LLM et ont tendance à se fier davantage aux explications plus longues (c'est-à-dire « biais de longueur »). L'incapacité des utilisateurs à discerner la fiabilité des réponses des LLM non seulement compromet l'utilité de ces modèles, mais présente également des risques dans les situations où la compréhension de la précision du modèle par l'utilisateur est essentielle.

[70] Jakesch M., Bhat A., Buschek D., Zalmanson L., Naaman M. (2023), *Co-Writing with Opinionated Language Models Affects Users' Views,* Actes de la conférence CHI 2023 sur les facteurs humains dans les systèmes informatiques, disponible à l'adresse : https://arxiv.org/abs/2302.00560

[71] Rogiers A., Noels S., Buyl M., De Bie T. (2024), *Persuasion with Large Language Models: a Survey,* prépublication arXiv arXiv:2411.06837, disponible à l'adresse : https://arxiv.org/abs/2411.06837

[72] Phénomène lié à la flagornerie (privilégier l'accord des utilisateurs plutôt que les réponses indépendantes) qui pose des risques pour la fiabilité des IA conversationnelles, et qui peut être évalué à l'aide de benchmarks dédiés. Voir Fanous A., Goldberg J., Agarwal A.A., Lin J., Zhou A.Y., Daneshjou R. et Koyejo O. (2025), *SycEval : Evaluating LLM Sycophancy.* ArXiv, abs/2502.08177, disponible à l'adresse : https://arxiv.org/abs/2502.08177. Voir le phénomène à grande échelle de la persuasion automatisée dans l'article suivant : Matz, Sandra C., J D Teeny, Sumer S. Vaid, H Peters, Gabriella M. Harari et M Cerf. (2024), *The potential of generative AI for personalized persuasion at scale,* Scientific Reports 14, disponible à l'adresse : https://www.nature.com/articles/s41598-024-53755-0

[73] Zeng D., Legaspi R. S., Sun Y., Dong X., Ikeda K., Spirtes P. et Zhang K., (avril 2024), *Counterfactual reasoning using predicted latent personality dimensions for optimizing persuasion outcome*, in International Conference on Persuasive Technology (pp. 287-300), Cham: Springer Nature Switzerland, disponible à l'adresse : https://arxiv.org/abs/2404.13792

[74] Kaffee L., Pistilli G., Jernite Y. (août 2025), *INTIMA : A Benchmark for Human-AI Companionship Behavior,* ArXiv, abs/2508.09998, disponible à l'adresse : https://arxiv.org/abs/2508.09998.

[75] Rogiers, et al. (2024) op.cit.

[76] Voir l'affaire américaine *Garcia c. Character Technologies Inc* (affaire dite « Setzer », dans laquelle un garçon de 14 ans a développé un fort attachement émotionnel à un personnage conçu par Character.ai sur le modèle d'un personnage fictif de Game of Thrones).

[77] Des expériences demandant à divers LLM d'imiter ou de conseiller les décisions des gens dans des dilemmes moraux réalistes démontrent que les décisions et les conseils des LLM sont systématiquement biaisés contre toute action, et que ce biais est plus fort que chez les humains. De plus, elles présentent des preuves suggérant que ces deux biais sont induits lors du réglage fin des LLM pour les applications de chatbot. Voir Cheung V., Maier M., Lieder F. (2025), *Large*

language models show amplified cognitive biases in moral decision-making, Proc. Natl. Acad. Sci. U.S.A. 122 (25) e2412015122, disponible à l'adresse : https://doi.org/10.1073/pnas.2412015122

[78] Les systèmes conversationnels d'IA peuvent agir comme des chambres d'écho car les LLM ont tendance à approuver les opinions de leurs utilisateurs. Voir à ce sujet une étude quantitative menée par Nehring J., Gabryszak A., Jürgens P., Burchardt A., Schaffer S., Spielkamp M. et Stark B. (2024), *Large Language Models Are Echo Chamber,* dans Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 10117-10123, Turin, Italie. ELRA et ICCL, disponible à l'adresse : https://aclanthology.org/2024.lrec-main.884/

[79] *Bradshaw et autres c. Royaume-Uni,* requête n° 15653/22, arrêt du 22 juillet 2025, p. 135.

[80] *Sanchez c. France,* requête n° 45581/15, arrêt du 15 mai 2023, p. 185.

[65] Voir Kosmyna N., Hauptmann E., Yuan Y.T., Situ J., Liao X., Beresnitzky A.V., Braunstein I., Maes P. (juin 2025), *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task,* prépublication arXiv arXiv:2506.08872, non encore évaluée par des pairs, disponible à l'adresse : https://arxiv.org/abs/2506.08872

[82] Voir les études empiriques sur les biais humains amplifiés et l'introduction de biais potentiellement problématiques par l'utilisation de LLM, notamment Cheung V., Maier M., Lieder F. (2025), *Large language models show amplified cognitive biases in moral decision-making*, Proc. Natl. Acad. Sci. U.S.A. 122 (25) e2412015122, disponible à l'adresse : https://doi.org/10.1073/pnas.2412015122

[83] Voir les expériences sur la persuasion latente dans Jackesh et Zeng D., Legaspi R. S., Sun Y., Dong X., Ikeda K., Spirtes P. et Zhang K., (avril 2024), *Counterfactual reasoning using predicted latent personality dimensions for optimizing persuasion outcome*, in International Conference on Persuasive Technology (pp. 287-300), Cham: Springer Nature Switzerland, disponible à l'adresse : https://arxiv.org/abs/2404.13792

[84] Voir les articles 5 et 7 de la Convention-cadre sur l'intelligence artificielle.

[85] Voir également Bai H., Voelkel J., Eichstaedt J., Willer R. (septembre 2023), *Artificial intelligence can persuade humans on political issues*, disponible à l'adresse suivante : https://www.nature.com/articles/s41467-025-61345-5 ; voir également Commission de Venise « Déclaration interprétative du Code de bonnes pratiques en matière électorale concernant les technologies numériques et l'intelligence artificielle », CDL-AD(2024)044, paragraphe 21 : « Des élections démocratiques ne sont pas possibles sans le respect, entre autres, de la liberté d'expression, y compris la liberté des médias. Toute restriction à ces droits doit avoir une base légale, être nécessaire et dans l'intérêt public, et respecter le principe de proportionnalité ».

[86] Pour un aperçu des normes juridiques pertinentes du Conseil de l'Europe, voir : https://www.coe.int/fr/web/children/legal-standards. Voir également CM/Rec(2018)7 du Conseil de l'Europe intitulée « Lignes directrices relatives au respect, a la protection et à la réalisation des droits de l'enfant dans l'environnement numérique » et le document de référence du Comité de Lanzarote intitulé « Menaces et opportunités des nouvelles technologies pour la protection des enfants contre l'exploitation et les abus sexuels ».

[87] Si elle n'est pas strictement régulée, l'IA générative peut : a) empêcher les utilisateurs de bien comprendre les notions de dignité humaine, de consentement et de respect mutuel, notamment ceux qui sont encore en train de développer leur perception de la nature du lien relationnel ; b) renforcer des dynamiques relationnelles irréalistes, toxiques ou dysfonctionnelles, y compris par la normalisation ou l'automatisation de la manipulation ou de la disponibilité constante ; c) affaiblir la capacité des utilisateurs à faire la distinction entre l'interaction artificielle et l'interaction humaine authentique ; et d) entraîner les utilisateurs dans des formes automatisées de cyberharcèlement, interférant ainsi avec le développement émotionnel et l'image de soi.

[88] Voir Cools H., Diakopoulos N. (2024), *Uses of generative AI in the newsroom: Mapping journalists' perceptions of perils and possibilities*, Journalism Practice, 1-19, ou des commentaires antérieurs sur l'adoption du journalisme https://charliebeckett.medium.com/what-we-have-learnt-about-generative-ai-and-journalism-and-how-to-use-it-7c8a9f5e86fd.

[89] Voir Lignes directrices sur la mise en œuvre responsable des systèmes d'intelligence artificielle (IA) dans le journalisme, adoptées par le Comité directeur sur les médias et la société de l'information (CDMSI) le 30 novembre 2023.

[90] van Drunen M. Z. (août 2025), *Safeguarding media freedom from infrastructural reliance on AI companies: The role of EU law,* Telecommunications Policy, volume 49, numéro 7, disponible à l'adresse suivante : https://www.sciencedirect.com/science/article/pii/S0308596125000874

[91] Voir, par exemple : API, gouvernement Islandais, disponible à l'adresse : https://openai.com/index/government-of-iceland/

[92] Des études ont montré que certains modèles linguistiques sont nettement plus susceptibles de créer des lettres de motivation au ton moins formel (par exemple en termes de structure syntaxique et de formulation) lorsqu'il s'agit de profils féminins comparés à des profils masculins. En outre, les choix lexicaux reflètent fréquemment des stéréotypes et des biais de genre. Voir Wan Y., Pu G., Sun J., Garimella A., Chang K. W., Peng N. (octobre 2023), *« Kelly is a warm person, Joseph is a role model » : Gender biases in LLM-generated reference letters,* prépublication arXiv arXiv:2310.09219, disponible à l'adresse : https://arxiv.org/search/cs?searchtype=author&query=Wan,+Y

[93] Campbell C.H. (2024), *Automated Journalism at the Intersection of Politics and Black Culture: The Battle against Digital Hegemony*, Lanham, Maryland: Rowman and Littlefield.

[94] Au sens large ; voir par exemple l'approche holistique mise en œuvre par le Media Pluralism Monitor selon quatre dimensions : (i) protection fondamentale (des droits fondamentaux à la liberté d'expression et à l'accès à l'information, statut et sécurité des journalistes), (ii) pluralité du marché (en tenant compte à la fois des marchés numériques et traditionnels, de la production, de la distribution et de la consommation de contenus), (iii) l'indépendance politique (d'une salle de rédaction, mais aussi d'une structure et de ressources médiatiques et informationnelles plus larges), iv) l'inclusion sociale (accès et représentation de divers groupes sociaux, en particulier ceux qui se trouvent dans des conditions vulnérables), disponible à l'adresse suivante : https://cmpf.eui.eu/media-pluralism-monitor/

[95] *Centro Europa 7 S.R.L. et Di Stefano c. Italie,* requête n° 38433/09, arrêt du 7 juin 2012, p. 134.

[96] Voir le rapport technique de l'Autorité britannique de la concurrence et des marchés sur les implications des modèles fondamentaux d'IA en matière de concurrence (16 avril 2024), disponible à l'adresse suivante : https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf ; Rapport de l'Autorité française de la concurrence et des marchés de 2023, disponible à l'adresse suivante : https://www.autoritedelaconcurrence.fr/fr/communiques-de-presse/intelligence-artificielle-generative-lautorite-rend-son-avis-sur-le

[97] Toute restriction de la liberté d'expression, y compris lorsqu'elle est mise en œuvre par l'intermédiaire de l'IA, doit faire l'objet d'une analyse rigoureuse portant sur la légalité, le but légitime, la nécessité et la proportionnalité.
[98] L'article 8 de la convention-cadre, intitulé « Transparence et contrôle », est directement lié à cette partie. Au paragraphe 63, le rapport explicatif explique qu'en ce qui concerne le contrôle : « dans le contexte des systèmes d'intelligence artificielle, fait référence à divers mécanismes, processus et cadres conçus pour surveiller, évaluer et orienter les activités menées dans le cadre du cycle de vie des systèmes d'intelligence artificielle. Il peut s'agir de cadres juridiques, politiques et réglementaires, de recommandations, de lignes directrices éthiques, de programmes d'audit et de certification, d'outils de détection et d'atténuation des biais. Il peut également s'agir d'organes et de comités de surveillance, d'autorités compétentes telles que les autorités de supervision sectorielles, les autorités de protection des données, et les organes de promotion de l'égalité et des droits de l'homme, les Institutions nationales des droits de l'homme (INDH) ou les agences de protection des consommateurs, de contrôle continu des capacités actuelles de développement et d'audit, de consultations et de participation du public, de cadres de gestion des risques et de l'impact et les cadres d'évaluation de l'impact sur les droits de l'homme, les normes techniques ainsi que les programmes d'éducation et de sensibilisation.»
[99] Voir la méthodologie HUDERIA adoptée par le Comité sur l'intelligence artificielle (CAI) du Conseil de l'Europe lors de sa 12e réunion plénière, les 26-28 novembre 2024.
[100] Voir en particulier : Recommandation générale No.1 du GREVIO sur la dimension numérique de la violence à l'égard des femmes et Protéger les femmes et les filles contre la violence à l'ère du numérique : la pertinence de la Convention d'Istanbul et de la Convention de Budapest sur la cybercriminalité pour la lutte contre la violence à l'égard des femmes en ligne et facilitée par la technologie (2021).
[101] Voir la Convention des Nations Unies relative aux droits de l'enfant, articles 13 (droit à la liberté d'expression), 16 (droit à la vie privée et familiale) et 17 (diversité des médias visant à promouvoir le bien-être moral, physique et mental).
[102] Le filtrage et la restriction de la génération de contenus (y compris les contournements visant à échapper aux politiques de modération, appelés « *jailbreaking* ») peuvent nuire à la liberté d'expression, mais aussi affecter d'autres droits humains. S'il est dans l'intérêt public de limiter les contenus préjudiciables sur le suicide, la promotion de l'automutilation, les troubles alimentaires, les discours haineux, le terrorisme, le sexisme, etc., il est moins facile d'en identifier et de prévenir les substituts linguistiques et les euphémismes qui peuvent permettre le contournement.
[103] Voir Sanford J. (août 2025), *Why AI companions and young people can make for a dangerous mix,* Stanford Medicine News Center, disponible à l'adresse suivante : https://med.stanford.edu/news/insights/2025/08/ai-chatbots-kids-teens-artificial-intelligence.html
[104] Voir CM/Rec(2022)12 of the Committee of Ministers to the member States on electoral communication and media coverage of election campaigns.
[105] L'utilisation des normes internationales pertinentes (telles que ISO, IEEE, CEN/CENELEC) peut aider à élaborer conjointement des normes sociotechniques essentielles pour tester et évaluer les outils et applications d'IA générative en termes d'impact sur la liberté d'expression.
[106] Voir par exemple, au niveau européen, l'infrastructure numérique européenne (European Digital Infrastracture - EDIC) pour la préservation de la diversité linguistique et culturelle en Europe et la promotion de l'excellence et du leadership technologiques, appelée ALT-EDIC (https://www.alt-edic.eu/fr/).
[107] Y compris le respect des lois et règlements en matière de confidentialité.
[108] En ce qui concerne les normes éthiques, voir les organismes de normalisation tels que l'IEEE et l'ISO, qui ont élaboré des normes traitant spécifiquement des préoccupations éthiques liées aux systèmes d'IA. En outre, le CEN-CENELEC élabore actuellement des normes exigeant un niveau élevé de protection des droits fondamentaux dans le cadre de la mise en œuvre de la Loi sur l'IA de l'Union Européenne.
[109] Voir Comité des Parties à la Convention du Conseil de l'Europe sur la protection des enfants contre l'exploitation et les abus sexuels « Déclaration sur la protection des enfants contre l'exploitation et les abus sexuels facilités par les technologies émergentes », 7 novembre 2024.
[110] Voir Guide sur l'article 13 de la Convention européenne des droits de l'homme – Droits à un recours effectif, Conseil de l'Europe/Cour européenne des droits de l'homme, dernière mise à jour : 28/02/2025 ; voir également le « Chapitre IV – Recours » de la Convention-cadre sur l'IA, en particulier l'article 14 : « 1. Chaque Partie adopte ou maintient, dans la mesure où des voies de recours sont requises par ses obligations internationales et conformément à son ordre juridique interne, des mesures garantissant la disponibilité de voies de recours accessibles et effectives contre les violations des droits de l'homme résultant des activités menées dans le cadre du cycle de vie des systèmes d'intelligence artificielle. 2. Afin de renforcer la portée du paragraphe 1 ci-dessus, chaque Partie adopte ou maintien des mesures, y compris: a) des mesures garantissant que des informations pertinentes concernant les systèmes d'intelligence artificielle susceptibles d'avoir une incidence significative sur les droits de l'homme et leur utilisation pertinente sont documentées, fournies aux organismes autorisés à avoir accès à ces informations et, si nécessaire et applicable, mises à la disposition des personnes concernées ou communiquées à ces dernières; b) des mesures garantissant que les informations visées à l'alinéa a sont suffisantes pour permettre aux personnes concernées de contester la ou les décisions prises par le biais de l'utilisation du système ou fondées en grande partie sur celle-ci, et si nécessaire et approprié, de contester l'utilisation du système; et c) une possibilité effective donnée aux personnes concernées de former un recours auprès des autorités compétentes. » et l'article 15 – Garanties procédurales : « Chaque Partie veille à ce que, lorsqu'un système d'intelligence artificielle a un impact significatif sur la jouissance des droits de l'homme, les personnes affectées par celui-ci disposent de garanties, de protections et de droits procéduraux effectifs, conformément au droit international et au droit interne applicables. »
[111] Voir CM/Rec(2024)2 sur la lutte contre l'utilisation des poursuites stratégiques contre la participation publique (SLAPP).
[112] Voir Convention 108+, article 12 – Sanctions et recours, et chapitre IV – Autorités de contrôle.