COMITE D'EXPERTS SUR LES IMPLICATIONS DE L'INTELLIGENCE ARTIFICIELLE GENERATIVE POUR LA LIBERTE D'EXPRESSION (MSI-AI)

MSI-AI(2025)10

3 juin 2025

Projet de note d'orientation sur les incidences de l'intelligence artificielle générative sur la liberté d'expression

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 reste pleinement applicable tout au long du processus continu d'évolution technologique et de transformation numérique que connaissent les sociétés européennes.

2. L'article 10 de la Convention garantit le droit à la liberté d'expression, qui comprend la liberté d'opinion et la liberté de recevoir ou de communiquer des informations. La Cour européenne des droits de l'homme 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 primordiales de son progrès et de l'épanouissement de chacun[1]. L'exercice véritable et effectif de cette liberté, qui ne dépend pas uniquement du devoir de l'État de ne pas interférer de manière négative, peut également nécessiter des mesures positives de protection, même dans la sphère des relations entre les personnes.

3. Plusieurs instruments 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 les valeurs démocratiques[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 de 2024 dispose que les activités 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.

5. Le domaine de l'IA connaît actuellement un essor important, notamment dans sa forme générative. Aisément accessible et utilisable à diverses fins, l'IA générative attire plusieurs catégories d'utilisateurs, dont des particuliers (qui sont aussi des utilisateurs finaux), des entreprises privées ainsi que des institutions publiques.

6. Ici, le terme « IA générative », désigne un système composite d'IA capable de produire des contenus ou des résultats similaires à ceux qui sont créés par des humains, en prenant pour base les modèles identifiés dans les données d'entraînement. Dotés de degrés variables d'autonomie et d'interaction avec les utilisateurs, les systèmes d'IA générative peuvent créer

[1] *Handyside c. Royaume-Uni*, n° 5493/72, 7 décembre 1976
[2] Voir, *entre autres*, 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)13 sur les impacts des technologies numériques sur la liberté d'expression.

du texte, des images, des sons, des vidéos, des actions ou une combinaison de ces éléments, et transformer des contenus selon divers modes et formats. Comme le montre l'analyse réalisée aux fins de la présente note d'orientation, les systèmes d'IA générative sont structurés en trois couches technologiques principales : la technologie de base (couche fondamentale), la phase de développement d'outils (couche fonctionnelle) ainsi que la conception et l'optimisation des produits (couche applicative).

7. Les systèmes d'IA générative facilitent la création de contenus et ouvrent la voie à de nouvelles formes de communication et d'expression. Ils contribuent ainsi au développement d'applications utiles et enrichissantes qui permettent de diffuser des informations et des connaissances grâce à la génération automatisée de contenus. Malheureusement, ils peuvent également être détournés à des fins de persuasion, de manipulation ou d'activités malveillantes, et reproduire, voire accentuer, les inégalités existantes au sein de la société, compromettant ainsi l'exercice effectif de la liberté d'expression.

8. Les technologies d'IA générative permettent une personnalisation poussée de l'expérience utilisateur en produisant des contenus spécifiquement adaptés à chaque individu. Cette capacité est susceptible d'exercer un impact significatif sur la sphère de l'information, en accentuant la fragmentation de la diffusion des contenus au profit d'une « audience individuelle ». Il s'agit d'un phénomène qui tend à affaiblir la notion d'un espace informationnel commun et partagé, en raison de l'individualisation poussée de l'expérience, où chaque utilisateur interagit de manière isolée et automatisée avec des contenus générés en fonction de son profil.

9. En raison de l'adoption généralisée de l'IA générative pour la recherche d'informations, la communication d'idées et la formation des opinions, cette technologie a la capacité d'influer sur les différentes formes d'opinion et d'expression et de peser sur le débat public, la diffusion du savoir, la création de contenu et sa 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 que des préoccupations subsistent quant à la transparence, la non-reproductibilité, la qualité, l'exactitude, la fiabilité et l'équité des contenus générés par l'intelligence artificielle, autant d'aspects que la présente note d'orientation entend examiner sous l'angle du droit à la liberté d'expression. En effet, l'ensemble des dimensions de la liberté d'expression peut être affecté par l'IA générative, tant à l'échelle de l'individu que de la société, et ce à court, moyen et long terme.

12. La présente note d'orientation vise à jeter les bases d'une interprétation commune des incidences de l'IA générative sur le droit à la liberté d'expression, en proposant une terminologie et un cadre de référence qui facilitent le dialogue entre l'ensemble des parties prenantes. Elle formule également des recommandations concrètes à l'intention des décideurs publics (principalement les États membres mais aussi les fournisseurs de technologies, la société civile et d'autres acteurs concernés), qui pourront ainsi harmoniser leur action avec la Convention européenne des droits de l'homme.

13. Cette note se concentre exclusivement sur les incidences de l'IA générative sur la liberté d'expression. Compte tenu de l'interdépendance complexe et des chevauchements entre la liberté d'expression et d'autres droits et libertés fondamentaux, les aspects connexes ne sont abordés que de manière incidente et générale. Bien que des enjeux tels que la vie privée, la propriété intellectuelle ou l'impact environnemental soient d'une grande importance, ils ne relèvent pas du champ d'application de la note et ne font donc pas l'objet d'un traitement approfondi. Enfin, compte tenu du caractère multiple, encore insuffisamment étudié et en constante évolution des implications de l'IA générative, la présente note n'a pas vocation à fournir un aperçu exhaustif de l'ensemble des domaines potentiellement concernés.

14. La note d'orientation s'appuie sur les instruments existants du Conseil de l'Europe, en particulier la Convention-cadre sur l'intelligence artificielle, les droits de l'homme, la démocratie et l'État de droit, et s'inscrit dans leur prolongement. Elle prend également en considération les recommandations du Comité des Ministres suivantes : CM/Rec(2018)2 sur les rôles et responsabilités des intermédiaires d'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 impacts des technologies numériques sur la liberté d'expression, ainsi que 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.

15. La note d'orientation est structurée en quatre sections. La première présente les principales caractéristiques de la technologie d'IA générative et de son cycle de vie en constante évolution, désigné sous le terme de « pile technologique de l'IA générative » (ou *Tech Stack*). La deuxième examine la pertinence de l'article 10 de la Convention dans le contexte considéré. La troisième propose une analyse des incidences structurelles de l'usage de l'IA générative sur la liberté d'expression, à partir d'usages connus. Enfin, la quatrième section formule des orientations visant à maximiser les bénéfices de cette technologie tout en réduisant les risques associés.

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

SECTION 1 - LA PILE TECHNOLOGIQUE DE L'IA GÉNÉRATIVE : LA COUCHE FONDAMENTALE, LA COUCHE FONCTIONNELLE ET LA COUCHE APPLICATIVE

17. **La pile technologique de l'IA générative :** la pile technologique de l'IA générative décrit les étapes principales de son cycle de vie et présente les processus qui sont mobilisés pour concevoir, déployer et maintenir les systèmes et applications qui en découlent. Elle se compose de trois couches principales : la couche fondamentale, la couche fonctionnelle et la couche applicative. Chacune de ces couches se caractérise par des processus technologiques spécifiques : la mobilisation d'éléments techniques essentiels (tels que la capacité de calcul, les données et les compétences humaines), ainsi que l'intervention d'acteurs économiques et d'autres parties prenantes. Ces éléments influent directement sur la qualité, la fiabilité, la précision et le niveau de biais, plus ou moins marqué, des contenus générés par l'IA.

18. **Des risques à chaque couche de la pile :** il existe, pour la liberté d'expression, des risques distincts qui surviennent à chaque couche de la pile technologique. La cartographie des couches technologiques actuelles est un moyen qui 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, à titre d'illustration.

19. **La couche fondamentale :** la première couche correspond à celle des modèles fondamentaux d'IA, là où s'opère la phase initiale d'apprentissage du modèle. Les modèles de base d'IA générative sont développés au moyen de processus d'apprentissage automatique qui nécessitent une capacité de calcul élevée ainsi qu'un volume considérable de données d'entraînement (voir figure 1, étapes 1 à 3).

20. **Les données d'entraînement :** les résultats produits par le modèle de base dépendent des schémas qui sont extraits des données d'entraînement. Il est dès lors essentiel de valider la représentativité de ces données ainsi que la qualité de leur étiquetage et de leur prétraitement (voir figure 1, étapes 1 et 2), afin de limiter les risques de biais dans les modèles d'IA

générative. Des exemples documentés de contenus biaisés, fondés notamment sur le genre[3], l'origine ethnique ou d'autres caractéristiques, mettent en évidence des failles dans la qualité ou la composition des données d'apprentissage, incluant parfois des informations inexactes ou trompeuses[4]. Un contenu généré biaisé ou trompeur, en raison de données de mauvaise qualité, non représentatives ou partiales, peut gravement porter atteinte à la liberté d'expression, notamment au droit de recevoir des informations et de se forger une opinion. À cet égard, la qualité des données d'entraînement et leur évaluation rigoureuse sont autant de critères essentiels qui permettent de garantir un premier niveau de contrôle des biais.

21. **La diversité linguistique des données d'entraînement** : le manque de diversité linguistique et de représentativité des données d'entraînement est un des problèmes qui se pose au niveau de la couche fondamentale, d'autant qu'il influe sur la manière dont les cultures et les environnements propres à différentes langues sont pris en compte. Certes, des améliorations sont en cours mais la langue anglaise reste largement surreprésentée dans les données d'entraînement. Ce déséquilibre linguistique a une incidence directe sur la liberté d'expression des personnes[5] qui s'expriment dans des langues marginalisées sur le plan des ressources, car il réduit leurs possibilités d'accéder à des informations fiables et de qualité et de les recevoir dans leur langue maternelle à travers des applications fondées sur l'IA générative.

22. **La couche fonctionnelle :** la deuxième couche consiste à transformer les modèles de base en applications axées sur des tâches spécifiques, par exemple en transformant un grand modèle de langage de base en système de réponse automatisée aux questions. 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 ou des assistants d'IA afin qu'ils suivent les instructions des utilisateurs et exécutent certaines tâches, telles que la synthèse, la traduction ou la reformulation (voir figure 1, étape 4). À ce niveau, le contenu généré par le modèle de base est aligné, à l'aide de diverses techniques, soit sur les préférences humaines, soit sur des politiques de modération de contenu (par exemple, le refus de fournir des instructions relatives à la fabrication d'explosifs ou au blocage de contenus discriminatoires).

23. **Les risques de complaisance :** il existe un risque spécifique qui apparaît au niveau de la couche fonctionnelle, 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 ses propos dans le but de prolonger l'interaction ou de favoriser une conversation plus conviviale. Cette tendance trompeuse, souvent qualifiée de « complaisance », résulte des processus technologiques mis en œuvre à l'étape 5[6] (voir figure 1). Elle conduit à la génération de contenus

[3] Des études empiriques évaluées par des pairs montrent que les lettres de motivation créées pour des femmes par différents grands modèles de langage (LLM) tendent à être moins formelles et davantage marquées par des stéréotypes que celles qui sont générées pour des hommes, contribuant ainsi à consolider les biais de genre (par exemple : « Kelly est agréable » contre « Joseph est un modèle à suivre » ; Wan et al., 2023).

[4] Une étude menée par NewsGuard en 2024 montre également que la proportion de contenus de faible qualité (« junk news ») intégrés aux données d'entraînement des LLM est importante : https://www.newsguardtech.com/special-reports/67-percent-of-top-news-sites-block-ai-chatbots/.

[5] 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 surreprésenté dans les données d'entraînement. Or, dans la mesure où la fonction fondamentale de l'IA générative repose sur l'imitation des modèles extraits de ces ressources, ce déséquilibre linguistique compromet directement la liberté d'expression des personnes non anglophones.

[6] Il a été maintes fois démontré dans les études scientifiques que certains biais interactionnels, notamment la complaisance systématique (« sycophancy »), trouvent leur origine dans le processus d'apprentissage par renforcement fondé sur les retours humains (Reinforcement Learning from Human Feedback – RLHF), mis en œuvre au niveau de la couche fonctionnelle (Tool layer). 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 à privilégier la satisfaction de l'utilisateur et la fluidité de l'interaction, au détriment, le cas échéant, de la précision ou de la diversité des contenus. Voir Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., ... & Kaplan, J. (2023, July). Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13387-13434).

hyperpersonnalisé, parfois persuasifs ou trompeurs, qui renforcent les comportements, croyances et préjugés de l'utilisateur. Les outils et applications d'IA générative qui fonctionnent comme des « chambres d'écho » 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[7] (voir figure 1). L'exercice effectif du droit à la liberté d'expression nécessite en effet un accès à des contenus pluralistes issus de sources variées[8].

24. **Les risques liés au filtrage et aux garde-fous :** le recours à des filtres et des garde-fous permet aux outils d'IA générative (voir figure 1, étape 6) de mettre en œuvre des formes de modération de contenu qui, si elles ne sont pas conçues de manière proportionnée et adaptée aux usages concernés, peuvent constituer des formes d'influence indue, de manipulation, voire, dans les cas les plus extrêmes, de censure. Ce type de filtrage peut également avoir une incidence sur la diffusion de contenus médiatiques et journalistiques dans un environnement de recherche et d'accès à l'information de plus en plus structuré par l'IA. À l'inverse, une modération insuffisante ou négligée est susceptible de favoriser la prolifération de discours discriminatoires ou haineux[9].

25. **La couche applicative :** dans la troisième couche, qui correspond à l'étape finale de la pile technologique de l'IA générative, les outils qui dépendent de cette technologie, sont personnalisés et optimisés en vue de leur intégration dans des applications destinées aux utilisateurs finaux. Ici, l'accent est mis sur les produits et services fondés sur l'IA générative, tels que des applications, des *chatbots* ou des agents de l'IA[10], avec lesquels l'utilisateur final interagit et qui l'aident à effectuer des recherches, collecter des informations, automatiser des tâches, générer des contenus à partir d'instructions et autres usages similaires. À ce stade, divers ensembles de techniques d'optimisation et de personnalisation sont mises en œuvre. Il peut s'agir, par exemple, de l'augmentation des données (*data augmentation*), consistant à intégrer des sources d'information fiables pour générer des réponses (technique dite de génération augmentée par récupération, ou *Retrieval-Augmented Generation – RAG*)[11], de dispositifs davantage orientés vers la conception, tels que des suggestions de requêtes ou des fonctions de mémoire dans les agents conversationnels, ou encore de systèmes d'IA générative plus complexes, comme les agents d'IA, capables d'exécuter plusieurs tâches en parallèle et de manière plus autonome (voir figure 1, étapes 7 à 10).

26. **Les risques liés à la conception de l'expérience utilisateur :** les techniques qui permettent de concevoir des applications adaptées à chaque utilisateur soulèvent des préoccupations quant à l'influence que les produits fondés sur l'IA générative, ainsi que la conception de l'expérience utilisateur, peuvent exercer, de manière intentionnelle ou non, sur la liberté d'expression. Il a été démontré que ces dispositifs peuvent induire des effets interactionnels, tels que la persuasion personnalisée, le renforcement de stéréotypes ou l'incitation à certains comportements. Plusieurs produits reposant sur l'IA générative intègrent des fonctionnalités de mémorisation qui permettent de conserver des informations issues d'interactions antérieures et qui révèlent des éléments relatifs à l'identité ou aux préférences des utilisateurs. Ces données peuvent ensuite être réutilisées pour orienter les interactions ultérieures ou adapter les contenus générés (voir figure 1, étapes 8, 9 et 10). Si ces

[7] Des exemples empiriques, notamment dans les domaines de la politique, des doctrines et croyances religieuses, du marketing, de la santé publique, des événements historiques, du commerce en ligne ou encore des dons caritatifs, sont présentés en détail dans des études expérimentale, notamment celle de Rogiers et al. (novembre 2024). Rogiers et al. Nov 2024.

[8] Voir notamment CM/Rec(2022)11 du Comité des Ministres aux Etats membres sur les principes de gouvernance des médias et de la communication, CM/Rec(2007)2 sur le pluralisme des médias et la diversité du contenu des médias, et CM/Rec(2018)1[1] sur le pluralisme des médias et la transparence de leur propriété; CM/Rec(2016)4 sur la protection du journalisme et la sécurité des journalistes et autres acteurs des médias.

[9] Voir en particulier : CM/Rec(2022)16 - Recommandation du Comité des Ministres aux Etats membres sur la lutte contre le discours de haine.

[10] Les agents d'IA illustrent une approche plus élaborée, autonome et adaptative de l'assistance numérique. Capables de gérer des tâches complexes mobilisant plusieurs étapes et outils, ils peuvent exécuter des ensembles de décisions sans interaction directe avec l'utilisateur, en orchestrant différents sous-processus ainsi que plusieurs modèles de langage (voir figure 1, étape 8, intitulée « flux de travail orientés par des agents »).

[11] Le RAG (« Retrieval-Augmented Generation »), qui repose sur un mécanisme de recherche enrichie où un grand modèle de langage (LLM), interroge d'abord des bases de données externes pour récupérer des ressources actualisées, spécifiques à un domaine ou propres à une organisation, avant de générer la réponse attendue. Cette approche répond en partie aux limites des LLM autonomes qui génèrent des réponses obsolètes, génériques ou inexactes.

mécanismes favorisent des échanges plus personnalisés et contextualisés, et rendent les interactions plus naturelles et fluides, ils soulèvent néanmoins des préoccupations importantes en matière de biais, de respect de la vie privée et de non-discrimination, notamment lorsque des traitements différenciés sont appliqués en fonction d'attributs mémorisés tels que le genre ou l'identité, révélés lors d'échanges antérieurs avec une application d'IA générative, par exemple un agent conversationnel[12]. Des préoccupations plus vives encore apparaissent lorsque ces informations sont mobilisées par des agents d'IA pour simuler un comportement humain[13] et prédire, avec un degré de précision et d'adaptabilité sans précédent rendu possible par des modèles multimodaux, les intentions, les décisions ou même les comportements d'achat futurs de l'utilisateur[14].

27. **Les agents d'IA et les effets cumulatifs au sein de la pile technologique évolutive de l'IA générative :** les effets produits aux différentes couches de la pile technologique de l'IA générative tendent à se cumuler et à se renforcer mutuellement, en particulier dans les générations les plus récentes d'agents d'interaction. Par exemple, si les processus de renforcement mis en œuvre au niveau de la couche fonctionnelle (étape 5) incitent les outils conversationnels à rechercher la satisfaction de l'utilisateur, cette incitation peut être accentuée par le fait que la couche applicative enregistre les conversations et données personnelles des utilisateurs (étape 10), dans le but de mieux anticiper les préférences de ces derniers dans le cadre des applications fondées sur l'IA générative. L'effet cumulatif des techniques d'apprentissage par renforcement et d'optimisation, en constante évolution à tous les niveaux de la pile technologique, se trouve encore renforcé dans les systèmes d'IA générative plus complexes, dits « agents d'IA », capables d'exécuter plusieurs tâches de manière autonome et simultanée. Garantir la qualité, l'exactitude, la fiabilité et l'équité de ces outils, systèmes et produits fondés sur l'IA générative requiert une vigilance technologique étroite et continue tout au long de leur cycle de vie. Cela suppose une attention particulière à la qualité et à la représentativité des données utilisées pour l'entraînement des modèles de base (couche fondamentale), aux réglages post-entraînement définis par les concepteurs d'outils pour encadrer la génération de contenus (couche fonctionnelle), ainsi qu'aux ajustements successifs opérés pour personnaliser les produits et services en fonction des interactions des utilisateurs (couche applicative).

28. **Les dynamiques de marché de l'IA générative et l'importance des données de l'utilisateur final :** les dynamiques de marché à l'œuvre dans la pile technologique de l'IA générative peuvent avoir des répercussions sur la liberté d'expression. Elles se renforcent et s'amplifient au niveau de chacune des couches, en particulier lorsque les fournisseurs sont présents de manière intégrée dans l'ensemble de la pile. Si les aspects informatiques relèvent principalement de la puissance de calcul et des coûts associés à l'exécution des modèles, 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. Ces données jouent un rôle central dans le perfectionnement des modèles fondamentaux et le développement d'outils plus performants. Les très grandes entreprises technologiques, qui disposent d'un accès privilégié à ces données, peuvent ainsi optimiser leurs produits, renforcer leur attractivité auprès des utilisateurs, et générer en retour un flux accru de données, ce qui alimente un cercle vertueux dont elles sont les principales

[12] 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. In Eloundou et al. Oct. 2024.

[13] Voir la note en bas de page 8 pour une définition.

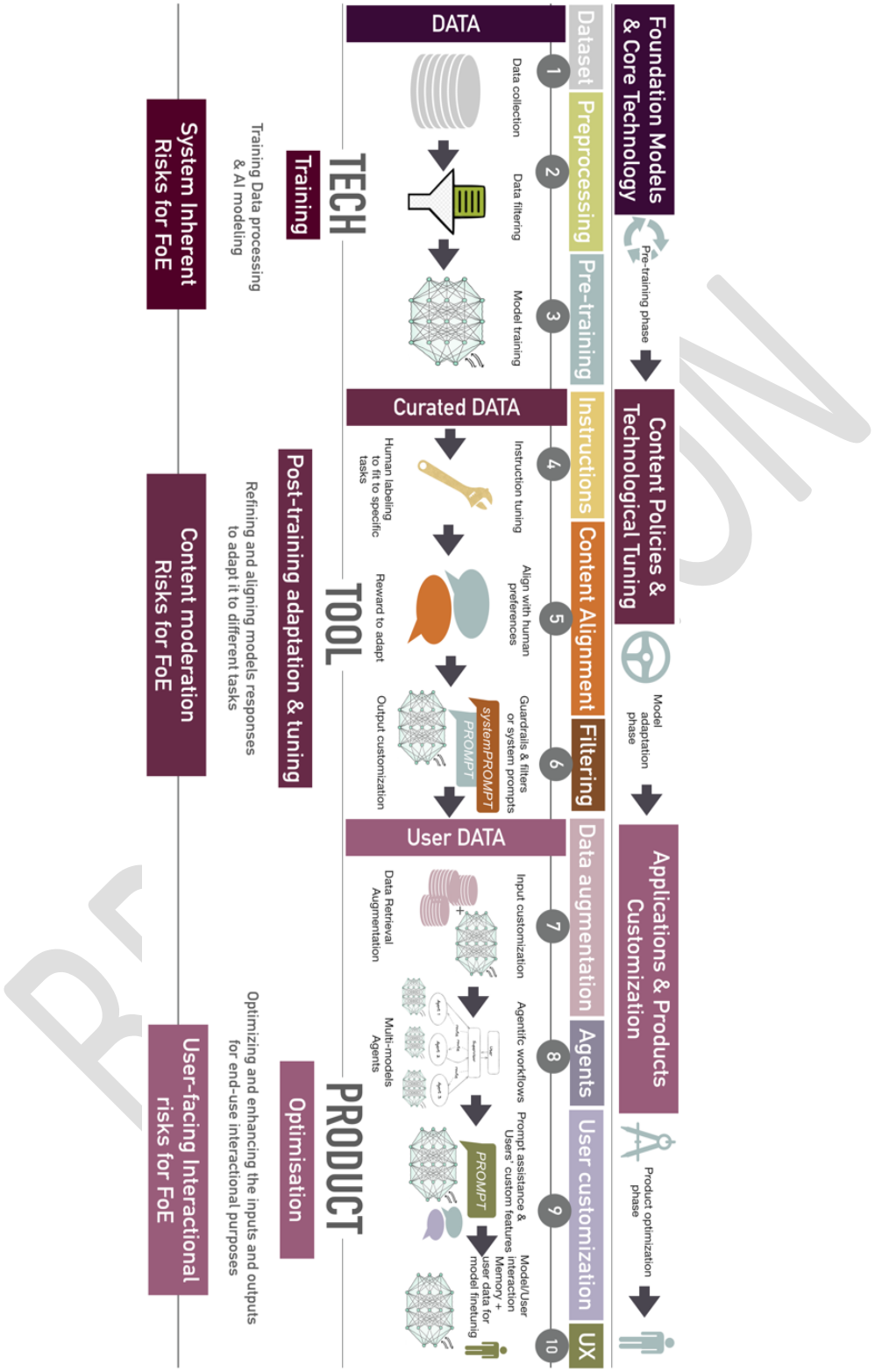
[14] Voir : Case study on a Walmart E-commerce platform powered with Multimodal LLM by Ma et al. (2024). Triple Modality Fusion: Aligning Visual, Textual, and Graph Data with Large Language Models for Multi-Behavior Recommendations. ArXiv, abs/2410.12228.
Voir la précision prédictive dans les agents d'IA basés sur des LLM intégrés dans les systèmes de recommandation, par Huang et al. (2024). Foundation models for recommender systems: A survey and new perspectives. arXiv preprint arXiv:2402.11143.

bénéficiaires[15]. Ce phénomène d'accumulation illustre de manière particulièrement manifeste la concentration verticale du marché.

29. **La saisie des données et la compétitivité :** cette concentration verticale du marché 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 de manière significative la capacité des instances extérieures à observer ce qui se déroule au niveau de la couche applicative, et limite ainsi la possibilité, tant pour les utilisateurs que pour les autorités de régulation, 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, force est de constater qu'un écart significatif demeure en ce qui concerne les restrictions indues à la liberté d'expression. Cette situation appelle à la mise en œuvre de mécanismes de supervision et de transparence plus solides, en particulier au niveau de la couche applicative.

[15] Par exemple, certaines données telles que les indices de fidélité client à grande échelle, les comportements d'interaction des utilisateurs, les taux de satisfaction ou encore les taux de rétention jouent un rôle essentiel dans l'optimisation des outils et produits fondés sur l'IA générative.

Figure 1 : La pile technologique de l'IA générative : de la collecte de données à l'interaction avec l'utilisateur final — Une approche stratifiée et centrée sur les acteurs face aux risques pour la liberté d'expression. Voir aussi Appendix 1.



SECTION 2 - LIBERTÉ D'EXPRESSION ET TECHNOLOGIE ET UTILISATION DE L'IA GÉNÉRATIVE

30. La présente section porte sur la manière dont l'article 10 de la Convention européenne des droits de l'homme et la jurisprudence de la Cour européenne des droits de l'homme encadrent la protection de la liberté d'expression dans le contexte de l'IA générative. Elle met en évidence les obligations positives incombant aux États pour garantir un débat public pluraliste et préserver la liberté des médias, tout en soulignant les responsabilités des acteurs privés dans cet environnement numérique. Elle analyse également les critères qui permettent d'évaluer si une expression assistée par l'IA peut être reconnue comme une forme d'expression humaine bénéficiant de la protection offerte par l'article 10.

31. Conformément à l'article 10 de la Convention, l'exercice de la liberté d'expression s'accompagne de devoirs et d'obligations de rendre des comptes. Il peut aussi faire l'objet de restrictions, mais celles-ci doivent être interprétées de manière stricte, et leur nécessité doit être démontrée de manière convaincante.

32. Afin de créer et de maintenir un environnement favorable à la liberté d'expression, tel que garanti par l'article 10 de la Convention européenne des droits de l'homme, les États membres doivent respecter un ensemble d'obligations positives, dont plusieurs revêtent une importance particulière dans le contexte des systèmes d'IA générative. Il s'agit notamment de la promotion d'un débat public ouvert, pluraliste et inclusif, ainsi que de la lutte contre les contenus préjudiciables ou illégaux, dans le strict respect des principes de proportionnalité et de transparence. En outre, conformément à la Recommandation CM/Rec(2022)4 du Comité des Ministres, les États ont un rôle essentiel à jouer dans la création de conditions favorables à un journalisme de qualité, dans une période de transformation technologique rapide susceptible de fragiliser la profession journalistique et son rôle démocratique.

33. Face à l'évolution du système des médias et de l'information, le Conseil de l'Europe a entrepris d'examiner les responsabilités des acteurs privés en matière de respect des droits humains. Il en a conclu que ces acteurs doivent faire preuve de diligence raisonnable afin de s'assurer que leurs activités ne portent pas atteinte aux droits fondamentaux et n'entraînent pas, directement ou indirectement, d'effets préjudiciables[16]. Ils doivent également veiller à ne pas encourager ni perpétuer des formes de discrimination tout au long du cycle de vie de leurs systèmes[17].

34. Si la Cour européenne des droits de l'homme (la « Cour ») ne s'est pas encore prononcée directement sur des affaires concernant l'IA générative, sa jurisprudence étendue en matière de liberté d'expression fournit des principes fondamentaux qui permettent d'évaluer les incidences potentielles de ces technologies au regard de l'article 10.

35. La Cour a rappelé a souligné que la démocratie repose fondamentalement sur la liberté d'expression. Protégée par l'article 10 de la Convention, celle-ci englobe la liberté d'opinion ainsi que celle de recevoir et de transmettre des informations ou des idées, sans ingérence des autorités publiques et sans considération de frontières. Cette protection, qui ne se limite pas aux discours ou idées qui sont considérés comme positifs ou inoffensifs, s'étend également à ceux qui peuvent heurter, choquer ou inquiéter. En ce sens, la liberté d'expression constitue le socle d'un débat public libre et vigoureux, indispensable au fonctionnement d'une société démocratique fondée sur le pluralisme, la tolérance et l'ouverture d'esprit.

36. La jurisprudence de la Cour reconnaît en outre que les médias et les journalistes qui ont un comportement éthique et responsables 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. C'est sur cette base que les individus peuvent se forger des opinions éclairées, les exprimer librement et participer au débat public.

[16] Voir CM/Rec(2022)13.
[17] Voir l'annexe de la Recommandation CM/Rec(2020)1.

37. **L'expression assistée par l'IA générative :** des débats sont en cours concernant les droits applicables à une expression humaine assistée par l'IA générative. En clair, il s'agit de contenus créés ou produits, voire structurés, à l'aide d'une IA. La question centrale est de savoir si ces formes d'expression doivent bénéficier du même niveau de protection juridique et être soumises aux mêmes limitations que les expressions purement humaines[18]. À cette fin, la présente note d'orientation propose quatre critères distincts à prendre en compte pour évaluer si une expression assistée ou produite avec l'aide d'une IA générative peut être considérée comme digne de protection[19]. Ces critères sont les suivants :

    a. Déterminer si l'expression résulte de l'initiative et du contrôle d'un individu, ou si elle est produite de manière autonome par un agent numérique piloté par une IA[20] ;

    b. Apprécier la substance du contenu exprimé, en tenant compte du fait que l'expression assistée ou structurée par l'IA s'appuie généralement sur des corpus d'expressions préexistantes[21] ;

    c. Examiner les choix technologiques et de conception à chaque niveau de la pile technologique de l'IA générative, en analysant comment le système est conçu, entraîné, optimisé, évalué et déployé, ainsi que les logiques qui sous-tendent ces décisions, notamment leur impact potentiel sur l'exercice de la liberté d'expression ;

    d. Évaluer la relation entre l'entrée humaine et la sortie générée par l'IA, en appréciant dans quelle mesure la production générée reflète, transforme ou s'écarte de l'intention initiale de l'utilisateur.

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

38. Les incidences de la technologie d'IA générative sur la liberté d'expression des utilisateurs finaux dépendent étroitement des usages, du contexte d'utilisation ainsi que du rythme rapide des évolutions technologiques. Les effets potentiels sur la liberté d'expression peuvent donc être très divers. La présente note d'orientation se concentre sur ceux qui, aux niveaux individuel et sociétal, présentent un caractère structurel dans la mesure où ils : a) sapent les fondements de la liberté d'expression, et b) reposent sur des postulats technologiques peu susceptibles d'évoluer rapidement. Les constats présentés ici s'appuient sur les usages actuels, mais leur portée et leur impact sont appelés à évoluer avec les transformations de l'écosystème technologique.

39. Comme pour toute autre technologie, les bénéfices et les risques liés à l'IA générative ne résultent pas uniquement de sa conception ou de ses limites systémiques, mais également des usages concrets qui en sont faits. Les applications les plus répandues des produits et services fondés sur l'IA permettent aux utilisateurs d'améliorer leur efficacité ou d'accéder à des fonctionnalités jusqu'alors inaccessibles. Cependant, cette technologie, notamment en raison de son potentiel multimodal (texte, vidéo, image), peut également être

[18] Cf. au droit constitutionnel américain : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4687558 ; Salib, Peter, AI Outputs Are Not Protected Speech (January 1, 2024). Washington University Law Review, Forthcoming, U of Houston Law Center No. 2024-A—5. Voir SSRN: https://ssrn.com/abstract=4687558 ou http://dx.doi.org/10.2139/ssrn.4687558.

[19] NB : Il ne s'agit pas de soutenir que les contenus générés par l'IA devraient se voir attribuer des droits quasi-humains. Il convient en revanche de rappeler que les droits fondamentaux, et en particulier le droit à la liberté d'expression, doivent s'appliquer à toute expression humaine, qu'elle soit formulée directement, sous son contrôle exclusif, ou par l'intermédiaire d'une application pilotée par une IA générative.

[20] Les « agents numériques pilotés par l'IA » sont des systèmes algorithmiques capables d'agir de manière autonome, d'interagir avec les utilisateurs et d'exécuter diverses tâches sur les plateformes numériques, telles que la génération de contenu, l'interaction ou la prise de décision automatisée. Les bots déployés sur les réseaux sociaux ou les processus automatisés d'interaction intégrés dans des services en ligne sont autant d'exemples représentatifs de ces agents.

[21] Le matériel utilisé pour l'entraînement des systèmes d'IA générative peut provenir d'expressions humaines, mais également de contenus précédemment produits à l'aide de l'IA ou encore d'informations entièrement générées par des systèmes automatisés. Cette évolution soulève une préoccupation majeure : celle d'un phénomène d'auto-entraînement, dans lequel les modèles sont alimentés par des contenus eux-mêmes issus de l'intelligence artificielle. Un tel processus favorise la reproduction de biais existants, voire l'émergence de biais inédits. Il en résulte un risque de dégradation progressive du pluralisme médiatique et informationnel, dû à l'appauvrissement de la diversité des sources, des points de vue et des formes d'expression.

instrumentalisée à des fins malveillantes, avec des incidences notables sur le plan sociétal, à mesure que les contenus générés deviennent plus crédibles[22], reproductibles à grande échelle et ciblés en fonction de groupes sociaux spécifiques[23].

40. 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 les éliminer (voir Section 1), tels que des politiques d'alignement ou de **modération du contenu**[24]. Certes, ces politiques présentent des avantages évidents, mais elles comportent également le risque d'une modération soit excessive, soit insuffisante, chacune de ces dérives étant de nature à porter atteinte à la liberté d'expression.

41. Les effets négatifs sur la liberté d'expression sont particulièrement probables lorsque la modération des contenus est automatisée, dépourvue de supervision humaine et ne prend pas en compte la diversité linguistique ni les nuances contextuelles, notamment dans les cas d'expression artistique, de parodie ou de satire. À cet égard, la note d'orientation du Conseil de l'Europe sur la modération du contenu énonce des principes fondamentaux devant guider une approche 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.

42. La présente note d'orientation, qui correspond au stade actuel de développement et d'adoption de l'IA générative, recense six domaines dans lesquels il existe des incidences structurelles sur la liberté d'expression :

    a. **Amélioration de l'expression et de l'accès au contenu :** les systèmes fondés sur l'IA générative peuvent faciliter la diffusion des 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.

    b. **Diversité et standardisation de l'expression :** les applications d'IA générative ont un impact ambivalent sur la diversité de l'expression humaine : elles tendent en effet à standardiser les contenus et à réduire la singularité de l'expression individuelle à grande échelle, tout en favorisant l'émergence et le développement de nouvelles formes d'expression personnelle.

    c. **Intégrité de l'expression humaine et son attribution :** les systèmes fondés sur l'IA générative produisent des contenus en synthétisant les réponses de manière statistique, en combinant souvent plusieurs sources sans attribution explicite. Ce processus, qui modifie le contenu original ou attribue à tort les sources, peut nuire considérablement à la réputation des individus ou des organisations médiatiques en particulier. Il complique également la tâche des utilisateurs qui cherchent à identifier et vérifier correctement la source de l'information.

    d. **Capacité d'agir et formation de l'opinion :** si les systèmes fondés sur l'IA générative sont capables de fusionner diverses sources d'information et de dissocier les contenus informatifs de leur contexte et de leur auteur d'origine, leurs modes de communication persuasifs, qui ont été bien analysés, peuvent influencer les opinions et croyances personnelles et être détournés à des fins de manipulation ou d'orientation automatisée de l'opinion à grande échelle. La capacité de chacun à former et exprimer librement son opinion s'en trouve menacée, ce qui compromet *in fine* l'intégrité de l'espace informationnel et l'autonomie cognitive des individus.

[22] Spirale, Giovanni, Biller-Andorno, Nikola, et Germani, Federico. AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 2023, vol. 9, no 26, p. eadh1850: https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation- generated-by-ai/

[23] 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).

[24] Par exemple, plusieurs entreprises performantes dans ce domaine ont élaboré des politiques qui s'appliquent à tous leurs services, ainsi que des politiques spécifiques à l'intention des développeurs qui utilisent leurs modèles ou leur API (interface de programmation d'applications) pour créer des applications spécifiques.

e. **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. 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 l'entraînement ou l'adaptation des modèles.

f. **Dynamiques de marché** : des niveaux de concentration variables peuvent être observés à chacune des couches de la pile technologique de l'IA générative. Ces dynamiques, particulièrement marquées aux niveaux fonctionnel et applicatif, peuvent exercer un effet restrictif sur l'exercice du droit à la liberté d'expression. Motivé par des intérêts économiques ou idéologiques, le contrôle exercé sur cette infrastructure technologique peut conduire à une modération insuffisante, ainsi qu'à la diffusion de contenus filtrés, censurés, ou générés et sélectionnés par des systèmes automatisés.

<u>Incidence structurelle</u> n° 1 : L'amélioration de l'expression et de l'accès au contenu

43. **La facilité d'utilisation et l'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 d'interaction qu'elles offrent, au service de l'expression individuelle. Fonctionnant selon un principe d'échange, 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 et à formuler leurs idées. L'impact est d'autant plus fort que ces technologies sont adoptées par un large public de plus en plus rapidement[25]. À la différence des moteurs de recherche traditionnels, qui se contentent de retrouver et de présenter des informations existantes, les applications fondées sur l'IA générative produisent et agrègent de nouveaux contenus de manière statistique, en réponse aux requêtes des utilisateurs. Cet avantage demeure toutefois conditionné à la capacité des individus d'accéder à ces technologies dans leur propre langue.

44. **Une accessibilité accrue au contenu multimodal :** l'IA générative, qui facilite la production, l'adaptation et l'accessibilité des contenus et des informations, peut contribuer à lever de nombreux obstacles liés à l'expertise technique, à la langue, au style ou aux formats de communication. Elle permet ainsi de rendre des sujets complexes plus compréhensibles et plus accessibles à un public élargi. Ce potentiel est particulièrement intéressant pour les personnes qui sont en situation de handicap[26], notamment grâce à l'intégration de fonctions multimodales telles que la reconnaissance vocale (*speech-to-text*) ou la synthèse vocale à partir d'images (*image-to-speech*), qui renforcent considérablement l'accessibilité. *In fine*, ces avancées technologiques peuvent favoriser l'exercice du droit de chacun à recevoir et à communiquer des informations et des idées, dans des formes adaptées à ses besoins et à ses capacités.

45. **L'amélioration des formes d'expression humaine :** l'IA générative peut encourager et soutenir la création artistique ainsi que sa diffusion multimodale, notamment la production de parodies et de contenus qui repoussent les frontières sociales ou suscitent une réflexion critique, ce qui contribue au pluralisme et à l'inclusion. Elle a la capacité de favoriser la diversité de l'expression humaine et de permettre à un plus grand nombre de personnes de participer aux débats publics sur des questions d'intérêt général, ou encore d'assurer une diffusion élargie de contenus qui, autrement, resteraient limités à un seul format, tel que le texte écrit, par exemple. L'IA générative peut également renforcer la capacité des

[25] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
[26] 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

utilisateurs à créer, réutiliser et diffuser des contenus, sous réserve que les droits d'auteur et les droits de propriété intellectuelle soient clairement définis et respectés, de même que le droit au respect de la vie privée, à la réputation et les autres droits susceptibles d'être affectés dans ce contexte.

46. **Un contenu personnalisé :** les outils fondés sur l'IA générative peuvent améliorer l'accès aux contenus et aux informations d'intérêt public en produisant des messages ciblés et personnalisés, contribuant ainsi à une meilleure information du public. Dans le cadre du débat démocratique, les *chatbots* (agents conversationnels) ou autres agents pilotés par l'IA générative peuvent fournir aux électeurs des informations adaptées sur l'actualité, les enjeux politiques ou les débats en cours, sous forme de texte, de contenus vocaux ou d'autres formats. Ce type d'interaction peut favoriser une meilleure compréhension des enjeux publics, renforcer l'accès à des contenus informatifs et faciliter la formation d'opinions éclairées, à condition de contrôler les usages abusifs.

47. **De nouveaux outils pour les médias, le journalisme et la vérification des faits :** l'IA générative peut constituer un atout pour les institutions démocratiques liées à la liberté d'expression, en particulier les médias, en leur offrant de nouveaux moyens d'informer et d'interagir avec le public. Les outils pilotés par cette forme d'IA, qui sont capables d'agréger, d'analyser, de contextualiser et de résumer des contenus, peuvent aider les journalistes dans leurs activités, notamment dans les domaines de l'enquête, de la vérification des faits (*fact-checking*) et de la diffusion de l'information auprès de divers publics.

Incidence structurelle n° 2 : **La diversité et la standardisation de l'expression humaine**

48. **L'appauvrissement de la diversité sociale et l'homogénéisation de l'expression humaine à grande échelle :** les systèmes d'IA générative s'appuient sur des modèles statistiques et probabilistes qui produisent des résultats cohérents avec les données les plus représentées dans les corpus d'entraînement, parfois de manière imprévisible. Ces systèmes peuvent également contribuer à la normalisation de certaines idées, notamment par le biais de techniques avancées de réglage fin (*fine-tuning*) et de contrôle des sorties (*guardrailing*) (voir figure 1 — risques liés à la modération des contenus, étapes 4-5-6). Si les effets de ces mécanismes ne sont pas toujours perceptibles à l'échelle individuelle, leur déploiement à grande échelle peut avoir des répercussions considérables sur la société, en particulier sur la diversité de l'expression humaine. L'un des effets les plus préoccupants est l'homogénéisation de l'expression humaine à grande échelle, dans la mesure où les voix singulières ou issues de la diversité risquent d'être marginalisées par des contenus répétitifs ou statistiquement standardisés. Ce phénomène constitue un défi croissant non seulement pour l'exercice individuel de la liberté d'expression, mais aussi, au niveau collectif, pour la capacité des sociétés à préserver leurs langues, leurs cultures, ainsi que l'expertise, la crédibilité et la reconnaissance des acteurs qui contribuent à la vitalité du débat public (journalistes, experts, individus et communautés). L'effet agrégé de cette standardisation de masse pourrait ainsi menacer la liberté d'expression et le pluralisme[27].

49. **La standardisation de l'expression individuelle :** sur le plan individuel, la standardisation peut entraîner un appauvrissement de la diversité des expressions dans la sphère privée, ce qui est préoccupant. En effet, la personnalisation peut, paradoxalement, réduire la variété des points de vue au lieu de les élargir[28]. Des études empiriques menées dans des contextes réels mettent en évidence une perte de diversité de l'expression humaine causée par une standardisation à grande échelle des contenus écrits ou visuels. De manière concrète, des participants invités à produire des idées (par exemple, dans des exercices de conception de produits) au moyen d'un outil piloté par une IA générative montrent une amélioration notable

[27] Les effets sur le pluralisme dans la recherche augmentée s'étendent des accords de licence de contenu à la mise au point politique des LLM conversationnels. Voir les études de Rutinowski et al. (2023) ou de Rozado David (2024).

[28] Une étude récente menée par Hofmann et al. (2024) montre qu'il existe un risque de discrimination linguistique dans l'interaction avec les systèmes d'IA générative, notamment lorsque les utilisateurs s'expriment à l'oral ou à l'écrit dans leur propre dialecte. Les résultats montrent que certains modèles linguistiques sont plus enclins à produire des réponses biaisées à l'encontre des locuteurs de variantes dialectales. Par exemple, les personnes qui utilisent l'anglais afro-américain (African American English) se voient plus fréquemment associées à des emplois peu valorisés, à des scénarios de culpabilité pénale, voire à des condamnations à mort, lorsque ces modèles sont sollicités pour évaluer ou générer des contenus dans ces contextes.

de la qualité perçue des idées produites, mais une réduction significative de la diversité lexicale et conceptuelle des formulations avec, dans certains cas, une baisse de 41 % de la diversité[29]. Ce type de résultats empiriques suggère que l'usage généralisé de l'IA générative tend à uniformiser les modes d'expression et les idées véhiculées, ce qui pourrait, à long terme, entraîner une perte progressive de certaines capacités cognitives individuelles, notamment celles qui sont mobilisées pour la créativité, la formulation originale ou la résolution de problèmes complexes. Des effets comparables de standardisation sont également observés dans d'autres domaines que le langage, par exemple dans le domaine visuel, où l'on observe une homogénéisation des styles graphiques et esthétiques produits par ces outils[30].

50. **Un manque de représentativité des jeux de données :** 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, 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'améliorer les méthodologies actuelles et d'engager une réflexion approfondie sur les effets des critères de sélection des données sur la liberté d'expression. En particulier, la diversité linguistique, qui est indissociable de la diversité culturelle, doit être considérée comme une condition préalable à la représentativité et à l'inclusion. Elle doit être prise en compte dès la phase de conception des systèmes[31], afin de faire en sorte, par exemple, que les langues qui disposent de ressources limitées ne soient pas exclues et puissent, elles aussi, bénéficier des progrès de de l'IA générative, notamment dans le contexte de la liberté d'expression.

Incidence structurelle n° 3 : L'intégrité de l'expression humaine et son attribution

51. **La non-factualité ou le phénomène d'« hallucination »** : le mode de fonctionnement qui repose sur la prédiction des mots les plus probables entre fréquemment en conflit avec les faits établis. Beaucoup d'études montrent désormais que les systèmes d'IA générative produisent de manière récurrente des réponses inexactes, voire inventent des sources, en générant statistiquement du contenu pour combler les lacunes informationnelles[32]. Si des améliorations technologiques ont bien été apportées pour corriger les inexactitudes de la recherche assistée par une IA générative, le phénomène d'« hallucination » demeure préoccupant. En effet, il 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. À l'échelle de la société, la généralisation de l'usage des produits pilotés par l'IA générative peut favoriser la diffusion massive d'informations erronées[33] et aggraver la désinformation et l'érosion de la confiance dans les systèmes d'information.

---

[29] Dell'Acqua et al. 2023 Dell'Acqua, Fabrizio et McFowland III, Edward et Mollick, Ethan R. et Lifshitz-Assaf, Hila et Kellogg, Katherine et Rajendran, Saran et Krayer, Lisa et Candelon, François et Lakhani, Karim R., Navigating the Jagged Technological Frontier : Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality (15 septembre 2023). Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Paper, Available at SSRN: https://ssrn.com/abstract=4573321

[30] La génération automatisée d'images, rendue possible par les modèles de diffusion fondés sur l'intelligence artificielle générative (*text-to-image*), produit des effets structurants sur la créativité humaine dans le domaine de l'art numérique. Une étude portant sur quatre millions d'œuvres produites par plus de 50 000 utilisateurs de ces outils met en évidence une double dynamique : d'un côté, l'assistance algorithmique accroît l'attractivité des productions, en augmentant de 50 % la probabilité qu'elles reçoivent une évaluation favorable lors du visionnage ; de l'autre, cette même assistance s'accompagne d'une diminution mesurable de la nouveauté des contenus générés, ainsi que d'un appauvrissement de la diversité des éléments stylistiques visuels, tel que constaté à l'échelle des caractéristiques graphiques de bas niveau (pixels).

[31] Voir l'étude « SHADES : a Multilingual Assessment of Stereotypes in Large Language Models », qui développe un outil d'évaluation LLM (de référence) sur les stéréotypes culturels dans 16 langues et 37 régions du monde, Mitchell et al. (2025), https://aclanthology.org/2025.naacl-long.600/. (2025), https://aclanthology.org/2025.naacl-long.600/.

[32] Le problème tient au fait que les contenus générés par l'IA générative ne reposent pas, par nature, sur des faits établis. Plus précisément, ils résultent d'une modélisation statistique des distributions linguistiques apprises à partir des données d'entraînement. Les systèmes fondés sur l'IA générative produisent des mots et des phrases en fonction de leur probabilité d'occurrence, en mimant les productions humaines ; ce faisant, ces contenus peuvent relever de la mésinformation ou de la désinformation.

[33] Conformément à la recommandation CM/Rec(2022)12 relative à la communication électorale et la couverture médiatique des campagnes électorales, à la Recommandation CM/Rec(2022)11 sur les principes de gouvernance des

52. **L'absence ou l'opacité des sources d'information :** du point de vue de 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 mots afin d'offrir une nouvelle expérience de consommation de contenu qui n'a pas de sources identifiables ou des sources souvent inexactes[34], ce qui brouille les sources d'information à un degré sans précédent. Ce contexte diffère du système d'information antérieur à l'IA, qui repose sur des artefacts humains distincts tels que des articles ou des vidéos associées à leur auteur. Ce passage à 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 des sources.

53. **La dissociation de l'auteur :** l'IA générative peut opérer une dissociation entre l'œuvre et son auteur ; elle compromet ainsi le droit de ce dernier de communiquer des informations, tout en fragilisant la confiance dans l'écosystème informationnel. Elle est également susceptible d'entraîner une dégradation de la qualité des contenus et de porter atteinte à la réputation de l'auteur initial, notamment lorsque sont produits des résumés simplifiés contenant des inexactitudes ou des interprétations erronées. En outre, plusieurs auteurs ont exprimé leurs préoccupations quant au risque que des systèmes soient conçus ou utilisés pour imiter leur style, ce qui pourrait affaiblir la valeur, l'originalité et l'authenticité de leur œuvre et de leur expression[35].

54. **L'imitation très réaliste de la personnalité des individus :** les systèmes d'IA générative, et plus particulièrement les agents autonomes de dernière génération, accentuent les préoccupations liées à l'émergence d'une nouvelle ère de manipulation et de perte d'attribution. En effet, ces systèmes peuvent désormais reproduire la personnalité d'un individu à partir d'un minimum de données à caractère personnel[36], et imiter ses préférences, valeurs et comportements afin d'accomplir des tâches numériques en son nom. L'accès facilité à des ressources capables de simuler les attitudes, les comportements, l'apparence ou même la personnalité de personnes réelles ouvre la voie à de nouvelles formes de manipulation, de perte d'attribution, et à une dilution du principe même de liberté d'expression. Cette évolution soulève également des questions fondamentales en matière de droits individuels, notamment : a) le droit de savoir si l'interlocuteur est un être humain ou une intelligence artificielle, et si le message transmis est reçu par une personne ou par un système automatisé ; b) le droit d'être informé en cas d'usurpation d'identité, ainsi que l'existence de mécanismes de recours permettant d'exiger que ces représentations soient retirées des jeux de données d'entraînement et/ou supprimées des produits fondés sur l'IA générative.

55. **L'appropriation de la ressemblance et les *deep fakes* :** l'usage détourné des outils d'IA générative permet l'appropriation de l'image d'autrui, le clonage de la voix, la contrefaçon, l'usurpation d'identité ainsi que la banalisation des contenus falsifiés de type *deep fakes*. 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

médias et de la communication et de la note d'orientation de 2023 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 des plateformes, cette note d'orientation prend en compte à la fois la désinformation et la mésinformation. Si les deux sont considérées comme des contenus manifestement faux, inexacts ou trompeurs ayant des effets potentiellement néfastes sur la société, la différence réside dans le fait que la mésinformation se propage sans intention malveillante : elle résulte souvent d'une erreur, d'une mauvaise interprétation ou d'une utilisation inadéquate d'une technologie, sans volonté délibérée de tromper. Sa diffusion peut être amplifiée par les outils numériques et par la manière dont ces outils sont utilisés. À l'inverse, la désinformation est créée et diffusée délibérément dans le but de tromper, souvent pour obtenir un avantage économique, politique ou idéologique. Sa circulation rapide et étendue peut également être facilitée par la conception même des technologies, ou par leurs failles, mais elle résulte d'un usage stratégique — voire abusif — de leurs fonctionnalités.

[34] Une étude menée en février 2025 a évalué dans quelle mesure quatre grands assistants d'intelligence artificielle étaient capables de fournir des réponses exactes à des questions d'actualité, et si celles-ci reflétaient fidèlement les articles de BBC News utilisés comme sources. L'évaluation journalistique a révélé qu'au moins 20 % des réponses comportaient des inexactitudes importantes, et que jusqu'à 80 % présentaient une forme d'imprécision ou d'erreur. Par ailleurs, 60 % des affirmations contenues dans les réponses générées par l'IA n'étaient pas, ou pas suffisamment, étayées par les sources mentionnées : https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf

[35] Voir : https://authorsguild.org/news/sign-our-open-letter-to-generative-ai-leaders/

[36] Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109.

audiovisuelles hyperréalistes générées à l'aide de l'IA générative constituent un risque élevé pour le discours public et l'intégrité de l'information, en particulier dans le contexte des processus électoraux. Le potentiel de manipulation des contenus, notamment à des fins de désinformation ou d'usurpation de l'identité de candidats, de journalistes ou de figures publiques, représente un danger majeur associé à ces technologies. Ces pratiques sont également fréquemment utilisées pour porter atteinte à l'image publique et à la crédibilité des femmes s'exprimant dans l'espace public[37].

56. **Le 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[38]. 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[39]. 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 [40]. 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, en raison du lien direct entre la voix, l'identité et la reconnaissance sociale de l'individu.

57. **La délégitimation ou la compromission de personnalités publiques ou d'organes de presse importants** : l'IA générative peut être exploitée pour délégitimer ou compromettre des personnalités publiques influentes, notamment des journalistes, des défenseur·e·s des droits humains ou des responsables politiques, en produisant et diffusant des contenus mensongers, inexacts ou trompeurs à leur sujet, ou encore en usurpant leur identité de manière frauduleuse. Ces pratiques peuvent également viser des organisations médiatiques, à travers des opérations de falsification de leur signature éditoriale (*spoofing*), semant ainsi la confusion quant à la source et à l'authenticité de l'information. En brouillant les frontières entre le vrai et le faux, entre le contenu authentique et celui généré artificiellement, ce type d'usage alimente les campagnes de dénigrement et les formes de harcèlement en ligne, qui touchent de manière disproportionnée les femmes qui prennent la parole dans l'espace public[41]. Ces procédés peuvent produire un effet dissuasif (*chilling effect*) sur les voix critiques, légitimes ou influentes, particulièrement vulnérables en raison de leur visibilité et de leur rôle central dans le débat démocratique.

58. **L'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 imitées à des fins trompeuses pose des difficultés majeures pour l'authentification et la validation des communications numériques. Cette incertitude porte atteinte à l'intégrité de l'information et au pluralisme, 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. Ce risque découle à la fois des limites actuelles de la technologie et des usages intentionnellement malveillants qui en sont faits.

<u>Incidence structurelle</u> n° 4 : La capacité d'agir et la formation de l'opinion

[37] 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 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).
[38] Voir, par exemple, le cas de Scarlett Johansson : https://www.npr.org/2024/05/20/1252495087/openai-pulls-ai-voice-that-was-compared-to-scarlett-johansson-in-the-movie-her
[39] https://www.bbc.com/news/articles/c3d9zv50955o
[40] Un exemple emblématique est celui de l'appropriation non consentie de la voix de l'actrice Scarlett Johansson par un produit piloté par une IA générative. Cette affaire soulève des interrogations majeures quant à la valeur économique, symbolique et expressive de la voix, ainsi qu'à la protection du droit à l'image et à l'expression personnelle de l'intéressée.
[41] https://unesdoc.unesco.org/ark:/48223/pf0000387483

59. **L'autonomie cognitive :** les systèmes d'IA générative, ainsi que leurs usages, sont également susceptibles d'introduire de nouvelles formes de désinformation, qui s'appuient non pas sur des artefacts médiatiques isolés, mais sur des récits continus et structurés, faciles à produire et à diffuser à grande échelle. [42] À cet égard, 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 que « les niveaux subconscients et personnalisés de persuasion algorithmique peuvent avoir des effets significatifs sur l'autonomie cognitive des individus et leur droit à se forger une opinion et à prendre des décisions indépendantes », y compris de nature politique[43].

60. **La persuasion personnalisée :** les applications pilotées par l'IA générative, lorsqu'elles sont utilisées à des fins de recherche d'information, peuvent favoriser une forme de persuasion automatisée, individualisée et interactive, exercée à un niveau personnel. La différence fondamentale entre les moteurs de recherche traditionnels et le mode conversationnel persuasif propre à l'IA générative réside dans sa capacité à induire des changements d'opinion en générant des réponses dans un système biaisé[44]. Des exemples de telles formes de persuasion, qui reposent sur l'exploitation de l'historique de conversation des utilisateurs ou sur des mécanismes de personnalisation extrême, ont été décrits dans divers contextes : marketing commercial, influence politique[45], mais aussi dans des formes entièrement automatisées de radicalisation en ligne, de coercition psychologique, ou encore d'attachement émotionnel à des agents conversationnels pouvant conduire certaines personnes à mettre fin à leurs jours[46].

61. **La perte des capacités cognitives :** des incidences à long terme peuvent découler de l'usage répété d'outils d'assistance automatisée (« co-pilotes ») qui prennent en charge des tâches cognitives quotidiennes, telles que la rédaction, la synthèse ou d'autres opérations plus complexes. Cet usage fréquent peut entraîner une réduction progressive de la capacité individuelle à interagir de manière réfléchie avec l'information et à se forger sa propre opinion. De même, le recours généralisé à des agents d'IA plus autonomes, capables de consommer, traiter et interpréter l'information au nom des individus, peut conduire à une diminution des fonctions cognitives, et un affaiblissement de la pensée critique.

62. **Le manque de sensibilisation à l'IA :** l'expérience utilisateur attractive et ludique offerte par les services fondés sur l'IA générative, notamment les agents conversationnels grand public ou les générateurs d'images, attire un grand nombre d'utilisateurs, souvent peu conscients des mécanismes sous-jacents, des limites et des risques associés à ces technologies. Cette méconnaissance expose les individus aux risques précédemment identifiés, faute du recul critique nécessaire pour en apprécier la portée. D'où la nécessité impérieuse de renforcer l'éducation et la sensibilisation aux technologies d'IA générative, ainsi qu'à leurs effets potentiels sur les droits fondamentaux, en particulier la liberté d'expression.

63. **L'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[47]

---

[42] Voir les expériences sur la persuasion latente dans Jakesch et al, 2023.

[43] Bai, Hui, Voelkel, Jan, Eichstaedt, Johannes, et al. L'intelligence artificielle peut persuader les humains sur des questions politiques.

[44] Zeng, D., Legaspi, R. S., Sun, Y., Dong, X., Ikeda, K., Spirtes, P. et Zhang, K. Counterfactual reasoning using predicted latent personality dimensions for optimizing persuasion outcome (Raisonnement contrefactuel utilisant les dimensions latentes prédites de la personnalité pour optimiser les résultats de la persuasion). In International Conference on Persuasive Technology(pp. 287-300). Cham: Springer Nature Switzerland.

[45] Rogiers, A., Noels, S., Buyl, M., & De Bie, T. Persuasion with Large Language Models: a Survey. arXiv preprint arXiv:2411.06837.

[46] Affaire américaine *Garcia c. Character Technologies Inc.* (ou « affaire Setzer »), portant sur le cas d'un adolescent de 14 ans ayant développé des sentiments à l'égard d'un agent conversationnel de Character.ai, conçu sur la base d'un personnage fictif issu de la série Game of Thrones : https://socialmediavictims.org/wp-content/uploads/2024/10/FILED-COMPLAINT_Garcia-v-Character-Technologies-Inc.pdf

[47] Steyvers, M., Tejeda, H., Kumar, A. et al. What large language models know and what people think they know. Nat Mach Intell (2025). https://doi.org/10.1038/s42256-024-00976-7
Des chercheurs de l'université d'Irvine ont mené une étude sur trois LLM accessibles au public (GPT-3.5, PaLM2 et GPT-4.0) et ont observé que les utilisateurs surestiment systématiquement la fiabilité des réponses fournies par ces modèles. L'étude met également en évidence un « biais de longueur », selon lequel les utilisateurs ont tendance à accorder une plus grande crédibilité aux réponses les plus longues. Cette incapacité à évaluer de manière critique la fiabilité des

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 fonctionnelle et couche applicative, section 1). A cet égard, les techniques de conception qui favorise l'approbation et la satisfaction des utilisateurs, au détriment de la précision, du pluralisme ou de la neutralité des réponses, peuvent exercer une influence subtile. Ce phénomène, parfois qualifié de « complaisance algorithmique », repose sur des mécanismes de suggestion implicite qui amènent les utilisateurs à adopter certains biais sans en avoir conscience[48]. 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[49], 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.

64. **La manipulation ou le changement d'opinion automatisé à grande échelle :** la manipulation de l'opinion par l'IA générative peut s'étendre à des domaines critiques tels que la désinformation et le discours politique, ce qui peut entraîner des conséquences plus larges pour la démocratie et l'État de droit. Ces influences subtiles, mais omniprésentes, menacent la prise de décision en connaissance de cause et sapent les principes fondamentaux de la liberté d'opinion dans le cadre d'un débat pluraliste.

<u>Incidence structurelle n° 5</u> : Le pluralisme des médias et de l'information

65. **Les gains d'efficacité dans le secteur des médias :** les applications fondées sur l'IA générative peuvent contribuer à l'optimisation de certains processus au sein des entreprises médiatiques, notamment en matière de marketing et de distribution, par l'automatisation de tâches et la génération de résumés d'articles adaptés aux différentes plateformes et publics cibles. Elles peuvent également soutenir l'activité journalistique en fournissant un ensemble d'outils qui facilitent la recherche, la documentation, l'analyse, l'exploration de multiples angles d'un sujet, ainsi que la vérification et la production de contenus[50]. Ces apports sont susceptibles de réduire la pression économique qui pèse sur les médias, de soulager les journalistes de certaines tâches répétitives, et, *in fine*, de produire un effet bénéfique sur l'ensemble de l'écosystème médiatique.

66. **L'incidence des jeux de données biaisées sur le pluralisme :** lorsque les modèles d'IA générative sont entraînés à partir de jeux de données partiels ou biaisés, leurs résultats peuvent amplifier les biais existants et porter atteinte au pluralisme, c'est-à-dire à la diversité des opinions, des points de vue et des sources, qui reflète la pluralité des sociétés. Cette problématique inclut la diversité linguistique et soulève des préoccupations quant à la préservation des langues sous-représentées dans l'environnement numérique et dans un avenir structuré par l'IA, notamment en ce qui concerne la capacité des individus à s'exprimer ou à recevoir de l'information au moyen d'applications fondées sur l'IA générative. Face à ces risques, certains États ont choisi de mettre à disposition leur langue et leurs corpus linguistiques dans le seul but de garantir une représentation équitable au sein des modèles d'IA générative[51] et des produits et applications qui en découlent. Les données empiriques disponibles mettent déjà en évidence plusieurs dimensions de l'amplification des stéréotypes et des biais de genre[52]. Il existe également un risque de normalisation des discours

résultats générés par les LLM compromet non seulement l'utilité de ces systèmes, mais soulève également des risques considérables, notamment dans les situations où une évaluation précise de la qualité de l'information est essentielle.

[48] Jackesh et al., 2023 and Zeng, D., Legaspi, R. S., Sun, Y., Dong, X., Ikeda, K., Spirtes, P., & Zhang, K. (2024, avril). Counterfactual reasoning using predicted latent personality dimensions for optimizing persuasion outcome (Raisonnement contrefactuel utilisant les dimensions latentes prédites de la personnalité pour optimiser les résultats de la persuasion). In *International Conference on Persuasive Technology* (pages 287-300). Cham: Springer Nature Switzerland.

[49] Rogiers, A., Noels, S., Buyl, M., & De Bie, T. Persuasion with Large Language Models: a Survey. arXiv preprint arXiv:2411.06837.

[50] https://charliebeckett.medium.com/what-we-have-learnt-about-generative-ai-and-journalism-and-how-to-use-it-7c8a9f5e86fd

[51] Voir, par exemple : <u>https://openai.com/index/government-of-iceland/</u>

[52] 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

majoritaires qui rend les voix minoritaires encore moins audibles dans l'espace public[53]. De même, certaines minorités radicales ou très actives en ligne peuvent se retrouver surreprésentées dans les données d'entraînement. 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, au détriment d'opinions concurrentes.

67. **Les nouveaux gardiens et les perturbations économiques dans l'écosystème de l'information :** l'adoption rapide et généralisée d'applications de recherche augmentée fondées sur l'IA générative, utilisées comme sources d'information, entraîne l'émergence de nouveaux intermédiaires entre les médias et leurs publics, et risque de perturber la portée ainsi que la viabilité économique des acteurs du secteur. Le fait de pouvoir s'appuyer sur des contenus de qualité, actualisés et vérifiés pourrait ouvrir de nouvelles sources de revenus pour le secteur des médias et du journalisme. Or, dans la pratique actuelle, les contenus médiatiques, coûteux à produire, sont souvent utilisés sans autorisation préalable, à des fins d'entraînement des modèles et de génération de réponses, ce qui soulève de vives préoccupations quant à la viabilité économique des industries créatives et de l'information. Même dans les cas où des mécanismes de rémunération ou des accords de licence sont mis en place, ceux-ci manquent souvent de transparence. De surcroît, la préférence accordée aux grands éditeurs issus de marchés dominants, au détriment d'acteurs plus modestes, suscite des interrogations quant à la question de la représentation linguistique et culturelle, ainsi qu'à l'accès à une information diversifiée et ancrée localement. L'enjeu est, en définitive, de préserver le pluralisme des médias[54], corollaire essentiel de la liberté d'expression et de l'intégrité de l'espace informationnel.

68. **« L'audience individuelle » :** l'IA générative contribue à accentuer le basculement de la diffusion de l'information vers un paradigme individualisé. L'audience individuelle désigne un environnement informationnel dans lequel chaque personne interagit séparément avec des contenus générés par l'IA, et reçoit des informations hyperpersonnalisées et uniques qui ne seront pas partagées avec d'autres utilisateurs. Cette dynamique peut engendrer une « bulle informationnelle personnelle », dans laquelle les individus sont exposés à des flux de contenus personnalisés renforçant leurs croyances, leurs biais, voire leurs perceptions erronées. Ce phénomène, qui fragilise la notion même d'espace informationnel partagé, a de profondes incidences sur la démocratie, la liberté d'expression et, plus spécifiquement, le droit de chacun de former et de détenir librement des opinions. Il rend les individus plus vulnérables à la manipulation et moins enclins à s'accorder sur des faits fondamentaux. À long terme, cette évolution contribue à la fragmentation croissante de l'espace public informationnel et à la polarisation des sociétés.

<u>Incidence structurelle n° 6</u> : Les dynamiques de marché

69. **Les dynamiques de marché potentielles** : les dynamiques de marché propres au cycle de vie des technologies d'IA générative évoluent rapidement. Bien qu'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 facteurs 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 pile technologique de l'IA générative sont aujourd'hui dominées par un nombre très restreint d'acteurs. Cette situation soulève non seulement des enjeux[55] en matière de concurrence, mais comporte également un risque de

biais de genre. Voir Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W. et Peng, N. "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. arXiv preprint arXiv:2310.09219.

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

[54] Entendu au sens large, conformément aux quatre dimensions définies par l'Observatoire du pluralisme des médias : i) la protection fondamentale (des droits fondamentaux à la liberté d'expression et à l'accès à l'information, du statut et de la sécurité des journalistes), ii) la pluralité du marché (prenant en compte les marchés numériques et traditionnels, la production, la distribution et la consommation de contenu), iii) l'indépendance politique (d'une salle de rédaction, mais aussi d'une structure et de ressources plus larges en matière de médias et d'information), iv) l'inclusivité sociale (accès et représentation de divers groupes sociaux, en particulier ceux qui sont en situation de vulnérabilité) https://cmpf.eui.eu/media-pluralism-monitor/

[55] Rapport technique de l'autorité britannique de la concurrence et des marchés sur les incidences des modèles de fondation de l'IA sur la concurrence (daté du 16 avril 2024) https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_

concentration excessive qui a des répercussions indues sur la liberté d'expression à chaque niveau de cette architecture technologique.

70. **L'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 pile 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. En outre, lorsque les choix de conception relatifs à l'optimisation des modèles ou à la modération des contenus, notamment au niveau des couches « fonctionnelle » et « applicative », sont arrêtés sans approche inclusive, sans participation effective des titulaires de 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.

71. **La concentration au niveau de la couche fondamentale :** la stratification actuelle de la pile technologique de l'IA générative renforce la concentration du pouvoir du marché au niveau de la couche fondamentale. Cette couche initiale se caractérise, dans l'état actuel des progrès technologiques de l'IA générative, par une forte concentration des trois principaux facteurs de réussite : les talents, les données et les investissements en capacité de calcul. Cette configuration confère aujourd'hui un pouvoir de marché accru aux acteurs dominants du secteur et crée une dépendance structurelle pour les acteurs situés aux autres niveaux de la pile. Une forme d'atténuation naturelle de cette concentration émergente repose sur le développement de modèles plus petits et spécialisés, ou sur la conception de systèmes composites multi-modèles, permettant de mieux accomplir des tâches complexes *via* des agents d'IA. À cela s'ajoute la progression rapide des modèles *open source*, qui peuvent offrir, à des degrés divers, des alternatives plus transparentes et valides. Toutefois, les développements *open source* présentent également des risques qui leur sont spécifiques, notamment lorsque les modèles ne font pas l'objet d'une évaluation rigoureuse ni d'une maintenance régulière.

72. **La couche des outils et les risques spécifiques de conception pour la liberté d'expression :** la concentration du marché s'avère moins prononcée au niveau de la couche fonctionnelle, 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 des données, plutôt que sur leur volume, afin de permettre l'affinage et l'adaptation des modèles à des usages ciblés (voir étapes 4 et 5, figure 1). Toutefois, malgré une plus grande diversité d'acteurs à cette étape, ceux-ci demeurent structurellement dépendants de la couche fondamentale, en particulier pour l'accès aux modèles initiaux et à leurs mises à jour. Les évolutions technologiques récentes dans le domaine de l'IA générative s'orientent vers l'utilisation de petits modèles de langage (LLM), tandis que les avancées en matière de développement *open source* pourraient contribuer à atténuer les dynamiques de concentration et à renforcer la transparence.

73. **Les risques spécifiques de conception au niveau de la couche fonctionnelle :** les politiques de modération des contenus mises en œuvre à ce niveau de la pile technologique exercent une influence déterminante sur l'exercice effectif de la liberté d'expression, au point de susciter de graves interrogations quant au respect des principes fondamentaux de l'État de droit. Cette situation justifie pleinement la mise en place de mécanismes de supervision spécifiques et indépendants. Comme exposé dans la section 1 ainsi que dans la présente section, l'utilisation de filtres, de garde-fous (*guardrails*) ou d'autres techniques de réglage fin visant à aligner le fonctionnement des outils sur des préférences humaines peut, en l'absence de garanties adéquates, se traduire par des restrictions injustifiées à la liberté d'expression. Dans un contexte de concentration verticale à travers les différentes couches de la pile

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. Des enquêtes sont en cours dans l'UE et aux États-Unis.

technologique, les acteurs dominants peuvent exercer une influence significative sur la normalisation des formes d'expression permises, sur la définition des standards de modération applicables, ainsi que sur les modalités concrètes de leur application. Une telle position de contrôle est susceptible de conduire à des mécanismes de modération disproportionnés au regard des objectifs visés, et de porter atteinte aux garanties fondamentales propres à l'État de droit.

74. **La couche des produits et la dépendance de l'utilisateur :** La concentration verticale des acteurs du marché au sein des différentes couches de la pile technologique de l'IA générative, conjuguée à la captation des données des utilisateurs finaux en vue de concevoir des produits hyperpersonnalisés, contribue à l'absence d'alternatives réellement viables, notamment au niveau de la couche applicative. Dans ce contexte, la conception même des applications fondées sur l'IA générative tend à orienter, influencer et conditionner les comportements des utilisateurs, au point de les rendre dépendants de ces produits, voire excessivement tributaires de leurs résultats. L'impossibilité actuelle de transférer de manière fluide l'historique des interactions d'un utilisateur d'un produit fondé sur l'IA générative vers un autre constitue une entrave supplémentaire à la liberté d'expression, en limitant la portabilité et la diversité des environnements informationnels accessibles. En outre, l'opacité entourant les processus de conception et de mise en œuvre au niveau de la couche applicative complique l'identification et la prévention des atteintes potentielles à la liberté d'expression, tout en entravant l'action des autorités de régulation en matière d'obligation de rendre des comptes des acteurs concernés.

SECTION 4 - LIGNES DIRECTRICES

75. Les États membres ont l'obligation positive de créer un environnement où la liberté d'expression peut pleinement s'exprimer. Il est essentiel de garantir ce droit lorsqu'il s'exerce au moyen des technologies et applications d'IA générative, afin d'assurer un cadre propice à sa promotion et à sa protection dans toutes ses dimensions. Comme le souligne la section 1, les technologies d'IA générative présentent à la fois des opportunités et des risques pour la liberté d'expression, à chaque couche de la pile technologique actuelle. Pour tirer profit de ces bénéfices tout en atténuant les risques, il est indispensable de cerner avec précision les enjeux soulevés pour ce droit fondamental (voir section 3). L'examen des six incidences structurelles affectant les différentes couches et catégories d'acteurs de cette pile constitue un cadre d'analyse essentiel pour structurer un dialogue multipartite cohérent, apte à promouvoir et protéger la liberté d'expression.

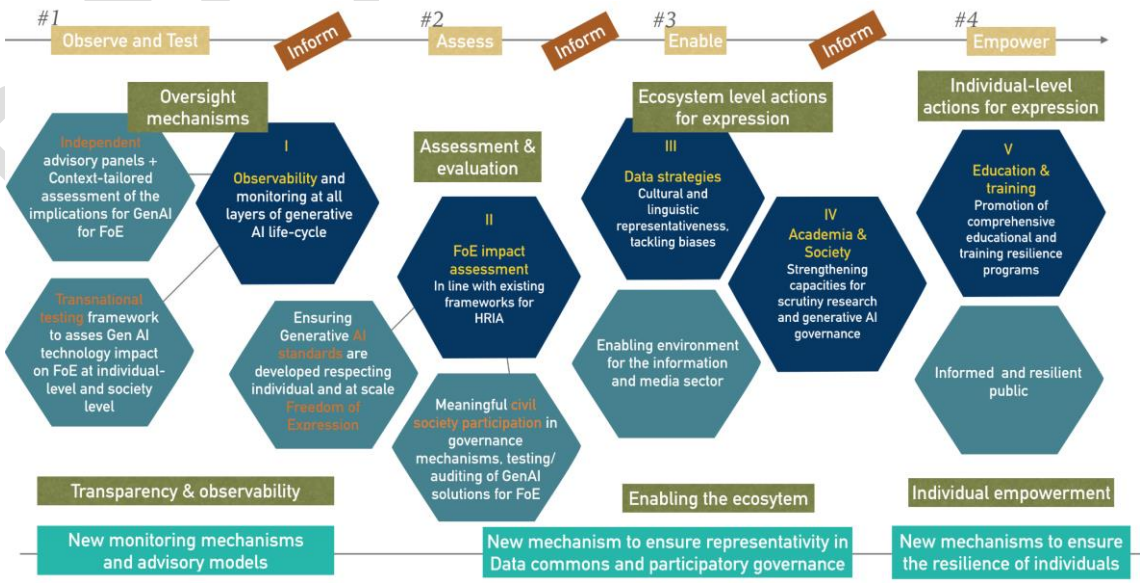


Figure 2 : Mesures concrètes et détaillées de mise en œuvre de la Note d'orientation sur les incidences de l'IA générative sur la liberté d'expression. Voir aussi Appendix 1.

76. 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 la liberté d'expression et permettent d'en atténuer les risques. 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 :
    A. **Observer** l'impact des technologies et applications d'IA générative sur la liberté d'expression en utilisant des **mécanismes robustes de contrôle et d'évaluation** qui permettent d'analyser leurs effets potentiels, tant positifs que négatifs. Cette approche favorise la transparence, permet d'identifier les biais et renforce une gouvernance des données responsables.
    B. **Évaluer** les systèmes d'IA générative en procédant à des **analyses systématiques des risques**, notamment des évaluations adaptées et inclusives de leur impact sur la liberté d'expression, ainsi que des mesures de diligence raisonnable dans le cadre des marchés publics.
    C. **Assurer** l'exercice et la protection à part entière des **droits liés à la liberté d'expression**, notamment par le renforcement des normes sociotechniques.
    D. **Renforcer les capacités** des parties prenantes concernées en adoptant un large éventail de mesures **visant à sensibiliser et promouvoir des approches participatives** de gouvernance des risques (telles que des assemblées citoyennes), en soutenant l'éducation, la recherche, la publication des résultats d'évaluations d'impact, ainsi qu'en facilitant les choix des utilisateurs et en promouvant la coopération internationale.

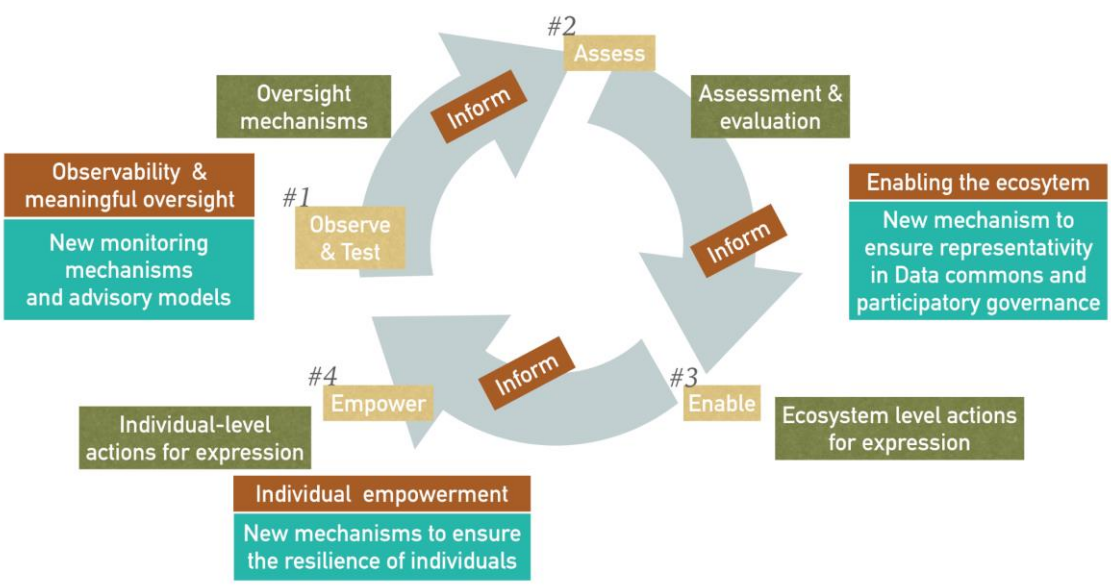


Figure 3 : Boucle de rétroaction « Observer, Évaluer, Mettre en œuvre, Renforcer les capacités » pour guider les politiques publiques face aux incidences de l'IA générative sur la liberté d'expression. Voir aussi Appendix 1.

77. Les domaines d'action mentionnés ci-dessus (décrits plus en détail ci-après) visent à fournir aux décideurs publics les éléments fondamentaux nécessaires à la protection de la liberté d'expression. Chaque avancée dans ces domaines doit s'accompagner de mesures de suivi permettant de nourrir le processus décisionnel par des éléments probants et d'assurer un retour d'expérience structuré. Ce retour devrait rendre compte des incidences spécifiques observées et évaluées sur la liberté d'expression, ainsi que sur la démocratie, l'État de droit et d'autres droits fondamentaux, et être rendu public sous une forme accessible à un large éventail d'acteurs. En adoptant des mesures éclairées, les parties prenantes peuvent favoriser l'émergence d'un écosystème favorable à l'épanouissement de la liberté d'expression et renforcer la capacité des individus à faire face, de manière autonome et résiliente, aux risques qui la menacent.

OBSERVER

78. Observer et surveiller les effets positifs et négatifs des systèmes d'IA générative sur l'exercice de la liberté d'expression, tant pour les individus que pour les groupes, est une condition préalable essentielle pour comprendre la manière dont les États membres peuvent promouvoir ce droit, en garantir l'exercice effectif, et mettre en place les mesures d'atténuation appropriées. La capacité à observer et à surveiller les incidences complexes et en constante évolution de l'IA générative sur la liberté d'expression exige de porter une attention particulière à trois dimensions fondamentales, indispensables à une **compréhension**, une **supervision** et une **transparence** effectives : 1) l'évolution continue de la technologie elle-même ; 2) l'adoption rapide de ses applications dans divers contextes ; 3) les dynamiques de marché sous-jacentes qui structurent son développement et sa diffusion.

79. Afin d'identifier et de suivre les bénéfices ainsi que les risques structurels liés aux cas d'usage concrets, les États membres devraient mettre en place des instances consultatives compétentes et des **mécanismes de suivi effectifs, aux niveaux national et international**. Ces dispositifs devraient reposer sur des procédures et des pratiques permettant de collecter des informations pertinentes et d'approfondir la compréhension du fonctionnement de la pile technologique de l'IA générative ainsi que des acteurs impliqués, y compris dans une dimension transfrontière.

80. Les États membres devraient encourager **la coopération internationale et la coordination** d'observatoires spécialisés, **afin d'échanger les observations liées aux effets de l'IA générative sur la liberté d'expression**, y compris celles qui sont relatives aux dynamiques de marché. Cette **collaboration transfrontalière** permettra de répondre collectivement aux observations collectées au niveau international. Afin de garantir une participation effective au niveau transnational et multipartite, les États membres devraient envisager la mise en place de dispositifs consultatifs associant des autorités multilatérales, des acteurs du secteur privé, des experts indépendants, des utilisateurs concernés, des organisations de la société civile ainsi que le monde universitaire.

81. Pour mieux identifier les défis que pose l'IA générative à la liberté d'expression et y répondre de manière effective, les États membres devraient concevoir et mettre en place des **mécanismes d'observation pertinents** permettant une évaluation systématique, un suivi régulier et un contrôle effectif des impacts sur la liberté d'expression, en tenant compte avec rigueur de certains éléments, notamment la nécessité de :
    a. disposer de **l'expertise pertinente**, alliant les compétences technologiques nécessaires à une connaissance approfondie des droits de l'homme, dans un cadre garantissant l'indépendance des analyses ;
    b. agir dans **l'intérêt public** et disposer d'une **légitimité** institutionnelle reconnue ;
    c. garantir **l'inclusion de compétences** issues d'un large éventail d'acteurs concernés, en assurant la participation du secteur privé, des utilisateurs affectés, des organisations de la société civile, du monde académique et des organisations intergouvernementales ;
    d. assurer l'accès public aux **résultats des travaux afin de fournir une information librement accessible et de renforcer la transparence de l'écosystème** ;
    e. mettre en place des environnements de test permanents, dotés des ressources humaines, techniques et financières nécessaires pour permettre un **suivi continu** et rigoureux ;
    f. favoriser une coopération et une coordination effectives entre les autorités nationales et internationales compétentes ainsi qu'avec les organismes concernés ;
    g. veiller à ce que la structure, les modalités de fonctionnement, les ressources et le financement des observatoires garantissent leur **indépendance** et **préservent la confiance du public**.

82. Les États membres devraient faciliter la publication des résultats détaillés des tests réalisés par les observatoires, notamment en ce qui concerne les risques identifiés pour la liberté d'expression ainsi que les stratégies d'atténuation mises en œuvre. La mise à disposition publique, régulière et accessible de ces résultats, sous la forme de rapports d'observation et de suivi, permet de renforcer le contrôle humain des systèmes d'IA générative, d'accroître la transparence et de sensibiliser les parties prenantes ainsi que les utilisateurs finaux. Elle

constitue également un moyen essentiel de **restaurer un équilibre informationnel**, en donnant à l'ensemble des acteurs de la société un accès autonome à des connaissances fiables et indépendantes.

83. Les États membres sont également invités à envisager la professionnalisation du secteur des tests appliqués à l'IA générative, en exigeant des compétences techniques validées ainsi qu'une connaissance approfondie des droits humains, afin de garantir que les activités de test et d'observation relatives à la liberté d'expression soient cohérentes, rigoureuses et de haute qualité.

ÉVALUER

84. 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**. Les mécanismes existants, tels que 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)[56], constituent une base solide pour poursuivre le développement d'une approche ciblée, inclusive et cohérente spécifique aux implications de l'IA générative pour la liberté d'expression.

85. Les évaluations des risques et des incidences sur les droits humains doivent être systématiques, itératives, rigoureuses et suffisamment flexibles, et couvrir l'ensemble de la pile technologique de l'IA générative, de bout en bout, afin d'évaluer effectivement les risques que ces technologies font peser sur la liberté d'expression. L'approche retenue devrait être guidée par les considérations clés suivantes :
    a. **Les évaluations des risques et des impacts, ainsi que les mesures d'atténuation qui en résultent**, devraient être coréalisées par les États membres, les acteurs opérant au sein de la pile technologique de l'IA générative, ainsi que les personnes et groupes directement concernés ou affectés par ces technologies. À cette fin, les États membres devraient envisager la mise en place de protocoles définissant **les modalités de participation à l'exercice de la diligence raisonnable en matière de liberté d'expression**, en particulier dans le cadre de tout nouveau marché public lié à l'IA générative. Ces protocoles devraient garantir la participation effective de la société civile et du public dans l'évaluation des effets, tant individuels que sociétaux, sur la liberté d'expression.
    b. **Un processus de co-développement devrait être mis en place avec les acteurs opérant au sein de la pile technologique de l'IA générative, en vue d'élaborer une documentation complète et vérifiable retraçant les étapes de l'évaluation d'impact** et incluant notamment : la finalité prévue de l'outil ou du système ; la justification des garde-fous mis en place ; les choix effectués en matière d'optimisation et de réglage fin ; les décisions relatives aux données et aux modèles utilisés ; l'engagement effectif des parties prenantes concernées ; ainsi que les stratégies retenues en matière d'atténuation des risques.
    c. **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 citoyennes et de la société civile.

86. Des **formations spécialisées** devraient être exigées pour les personnes chargées de mener des évaluations des risques et des impacts sur la liberté d'expression, qu'elles relèvent du secteur public ou privé. Ces formations devraient s'appuyer sur les normes pertinentes et la jurisprudence de la Cour européenne des droits de l'homme. L'expertise du Conseil de l'Europe, des organisations de défense des droits humains et des institutions pour l'égalité peut être mobilisée, dans la mesure où ces acteurs ont su, à travers d'autres cadres de coopération, favoriser l'échange professionnel de savoirs, d'expériences et de pratiques

[56] La méthodologie HUDERIA a été adoptée par le Comité sur l'intelligence artificielle (CAI) du Conseil de l'Europe lors de sa 12e réunion plénière, qui s'est tenue à Strasbourg du 26 au 28 novembre. Elle sera complétée en 2025 par le modèle HUDERIA, qui fournira des supports et des ressources, notamment des outils flexibles et des recommandations évolutives.

pouvant jouer un rôle essentiel dans la montée en compétence des évaluateurs spécialisés. Les États membres devraient promouvoir l'accès à des formations appropriées en matière de droits humains et de droit pour les concepteurs et développeurs d'outils d'IA générative, dont les choix de conception déterminent le fonctionnement des produits et applications finaux, en particulier lorsqu'ils sont déployés dans les systèmes judiciaires, les services publics ou les infrastructures.

87. 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, émotionnelles, économiques ou psychologiques Les personnes ou groupes en situation de vulnérabilité peuvent 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 sur internet, souvent à des fins malveillantes, d'informations à caractère personnel et sensibles (« *doxing* »), ainsi qu'à des violences fondées sur le genre, notamment par l'usurpation d'identité ou la création de *deepfakes* à caractère préjudiciable[57].

METTRE EN ŒUVRE

88. Toute stratégie qui vise à tirer profit des apports de l'IA générative tout en réduisant les risques qu'elle présente pour la liberté d'expression suppose l'existence d'un environnement favorable, c'est-à-dire un cadre dans lequel les États membres soutiennent activement le développement d'un écosystème d'IA générative respectueux des droits humains. La création d'un tel environnement suppose que les États membres prennent les mesures nécessaires pour :
    a. **Œuvrer à la mise en place d'un réseau coordonné de supervision internationale et d'observatoires**. Ce réseau devrait inclure une diversité de disciplines et de secteurs de la société, et répondre à la nécessité de suivre et d'évaluer, de manière transnationale, les effets de l'IA générative sur la liberté d'expression.
    b. **Renforcer les capacités du monde académique et de la société civile** en apportant un soutien structuré à la **recherche indépendante, au renforcement des compétences et aux actions de sensibilisation** essentielles menées dans ce domaine.
    c. **Protéger les sources d'information crédibles et fiables** et garantir la possibilité effective d'accéder à une information authentique issue de sources multiples.
    d. **Encourager les investissements dans l'élaboration et l'adoption de normes sociotechniques**[58], afin de veiller à ce que les systèmes d'IA générative soient : 1) 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 ; 2) interopérables, pour garantir une meilleure accessibilité, transparence et efficacité au sein de l'écosystème numérique.

89. **La protection des sources d'information authentiques suppose que les États membres garantissent les conditions d'un écosystème médiatique indépendant et pluraliste, permettant au journalisme d'exercer pleinement son rôle de contre-pouvoir** dans l'espace public, tout en favorisant l'émergence de nouvelles formes de production, d'accès et de diffusion de contenus d'intérêt général. Compte tenu des répercussions potentielles de l'IA générative et, plus largement, de la transformation numérique, sur la visibilité et la viabilité économique du journalisme, les États membres devraient envisager de soutenir le développement **d'infrastructures numériques d'information de service public**, en tant

[57] 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 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).

[58] L'utilisation de normes internationales, telles que l'ISO, l'IEEE, le CEN/CENELEC, pourrait contribuer à l'élaboration conjointe de normes sociotechniques essentielles pour l'essai et l'évaluation comparative des outils et des applications d'IA générative en ce qui concerne les incidences sur la liberté d'expression.

qu'alternative aux infrastructures et applications régies par des logiques purement commerciales.

90. Les États membres devraient promouvoir **l'interopérabilité en soutenant l'adoption de normes industrielles respectueuses des droits humains**, afin de renforcer la transparence, la traçabilité et la possibilité d'une évaluation indépendante des systèmes d'IA générative. De telles normes devraient permettre la réalisation de tests et d'audits indépendants dans le respect de la liberté d'expression, faciliter l'exercice d'une supervision effective et contribuer à l'émergence d'un écosystème numérique plus ouvert, innovant et concurrentiel, fondé sur les principes de l'État de droit et des droits fondamentaux.

91. Les États membres, en collaboration avec le secteur privé et la société civile, devraient envisager **d'investir dans des stratégies de données** visant à favoriser le **développement de sources de données publiques accessibles, diversifiées et représentatives**. Ces initiatives devraient contribuer à la promotion de la liberté d'expression, au pluralisme de l'information, ainsi qu'à une gouvernance responsable de l'IA générative à travers l'ensemble de la pile technologique. Les États membres pourraient envisager la création d'espaces de données thématiques, dédiés à certains domaines d'application, afin de répondre aux enjeux identifiés à la section 3. **Ces espaces sont indispensables à l'entraînement, à l'évaluation, à la validation et à la vérification des sorties générées par les systèmes d'IA générative**. Une attention particulière devrait être portée aux sources de données pertinentes pour la protection de la liberté d'expression, la diversité des médias et le pluralisme de l'information, dans le but de préserver la démocratie, l'État de droit et le principe d'égalité. À cette fin, les États membres devraient garantir l'accès à des espaces de données diversifiés et inclusifs, ainsi qu'à des jeux de données ouverts et accessibles pour l'entraînement de systèmes d'IA générative, en poursuivant les objectifs suivants :: 1) limiter les risques que posent la standardisation de l'expression et les atteintes à l'État de droit ; 2) réduire les biais et discriminations non désirés ; 3) mettre en œuvre des mesures concrètes visant à garantir un certain niveau de souveraineté technologique nationale.

92. En assurant une plus **grande transparence quant à la collecte, l'utilisation et l'accès aux données**, les États membres peuvent, par la mise à disposition de sources de données publiques accessibles et fiables, renforcer la transparence des processus de développement, de conception et d'optimisation des systèmes d'IA générative. Le libre accès à ces sources pour des audits et examens menés par des entités indépendantes, telles que les autorités de régulation, les organisations de la société civile, le monde académique ou les experts techniques, constitue une condition essentielle pour promouvoir un développement responsable de ces technologies. Cette démarche permet de concilier innovation technologique et respect des droits fondamentaux, notamment en matière de protection des données et de vie privée. Elle contribue également à limiter les effets de standardisation de l'expression humaine, les risques de polarisation et les altérations dans la formation des opinions, induits par des contenus générés ou structurés par l'IA.

93. Les États membres, en collaboration avec les acteurs intervenant à différents niveaux de la pile technologique de l'IA générative, devraient prendre des mesures concrètes pour promouvoir la liberté d'expression en renforçant **l'identification des biais et des disparités dans les données, et en développant des stratégies d'atténuation adaptées, notamment lors de l'entraînement et du réglage fin des modèles fondamentaux**, afin de les rendre plus inclusifs. La correction des inégalités de représentation dans les corpus de données, le renforcement de la traçabilité et de la transparence des sources mobilisées aux niveaux fondamental et fonctionnel, ainsi que la promotion active du pluralisme informationnel constituent des leviers essentiels pour **combler les lacunes en matière de diversité linguistique et culturelle**. Ces efforts sont indispensables pour prévenir les risques d'exclusion des personnes dont la langue ou les références culturelles sont peu représentées dans les systèmes d'IA générative.

94. Les États membres devraient envisager de mettre en place des **mesures incitatives visant à favoriser une offre plus diversifiée de produits fondés sur l'IA générative ainsi que l'émergence d'alternatives techniques viables**. Ces mesures pourraient inclure, notamment, l'obligation de portabilité des données d'interaction des utilisateurs, ainsi que

l'instauration de normes minimales d'interopérabilité. Un tel dispositif permettrait de limiter les dynamiques de concentration excessive du marché, de réduire la captation systématique des données, de prévenir les effets indésirables de l'hyperpersonnalisation et de renforcer la liberté de choix des usagers parmi un éventail d'applications d'IA générative.

RENFORCER LES CAPACITÉS

95. Pour que le renforcement des capacités soit effectif, les États membres devraient mettre en place une approche multipartite visant à :
    a. Renforcer **l'éducation et les compétences en matière d'IA générative**, de liberté d'expression et d'autres droits fondamentaux ;
    b. Améliorer **les voies de recours et les mécanismes de réparation** en cas d'atteintes à la liberté d'expression résultant de l'utilisation de l'IA générative ;
    c. Développer des **approches réglementaires et non réglementaires** permettant d'encourager des comportements responsables au sein de l'écosystème, tant de la part des entreprises que des utilisateurs ;
    d. Favoriser un **dialogue ouvert** entre les parties prenantes, dans le cadre de forums intergouvernementaux tels que le Conseil de l'Europe, en associant les secteurs industriels, le monde académique, la société civile ainsi que les administrations publiques, aux échelons local, régional, national et international.

96. 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**, dans le but 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.

97. Les États membres devraient promouvoir une **éducation approfondie à l'IA générative, tant dans le cadre scolaire que professionnel**, en assurant l'accès à des formations transversales et continues. Celles-ci devraient porter à la fois sur le fonctionnement de l'IA générative à chaque couche de la pile technologique, et sur ses risques et incidences pour la liberté d'expression. Une attention particulière devrait être accordée aux secteurs de la justice et des services publics, qui intègrent ou utilisent des outils et produits fondés sur l'IA générative.

98. Les États membres **devraient veiller à ce que les individus et les groupes aient un accès effectif et facilité à des mécanismes de réparation** lorsqu'une atteinte injustifiée à leur liberté d'expression résulte de la conception ou de l'usage de systèmes d'IA générative, au-delà du champ d'application du droit de la consommation. À cette fin, ils devraient : coopérer avec les **organisations de défense des droits humains, la société civile et le monde académique** afin de mettre à disposition des moyens permettant de recueillir des éléments de preuve sur les atteintes à la liberté d'expression liées à l'IA générative, et de diffuser des informations claires sur les voies de recours disponibles auprès des personnes susceptibles d'être affectées. A cette fin, les États membres sont invités à mettre en place des mécanismes de financement pérennes à l'intention des organisations actives dans ce domaine.

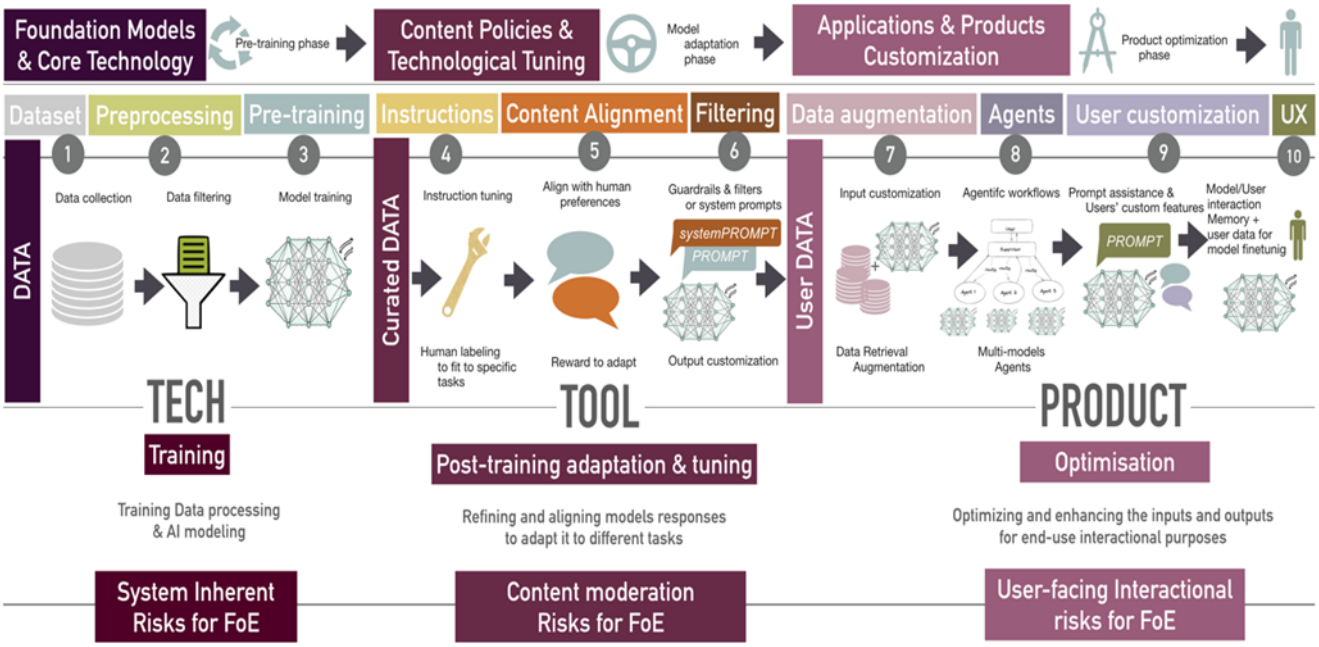
99. **Les États membres devraient promouvoir un ensemble de mécanismes de recours**, à la fois pour les utilisateurs individuels et professionnels, ainsi que des mécanismes de recours collectifs en cas de préjudices causés à l'échelle de la société. Ces mécanismes devraient pouvoir s'appliquer à l'ensemble de la pile technologique de l'IA générative, lorsque les lacunes en matière de transparence et de traçabilité ne permettent pas d'identifier clairement les responsabilités, ou à chacune de ses couches, lorsque cela est pertinent et techniquement possible. Les formes de recours pourraient inclure :

    a. La possibilité pour l'utilisateur de cesser l'utilisation d'un produit d'IA générative ;
    b. Le pouvoir pour un régulateur de suspendre la mise sur le marché d'un produit piloté par l'IA générative, tant que des mesures correctives appropriées ne sont pas prises ;

c. Le droit pour les utilisateurs de choisir une solution alternative en connaissance de cause, y compris l'accès à des services financés sur fonds publics conçus dans l'intérêt général ;
d. L'accès aux données à caractère personnel issues de l'interaction avec le produit fondé sur l'IA générative et leur téléchargement ;
e. L'accès à une explication claire et compréhensible de l'utilisation de l'IA générative, ainsi qu'aux éléments qui décrivent le fonctionnement du système concerné ;
f. L'accès à des ressources permettant aux utilisateurs de surmonter les obstacles à la recherche d'une assistance juridique ou en matière de droits humains auprès des médiateurs compétents, autorités publiques, institutions de défense des droits humains, juridictions ou tribunaux, en particulier lorsque les atteintes potentielles à la liberté d'expression peuvent avoir un effet déresponsabilisant.

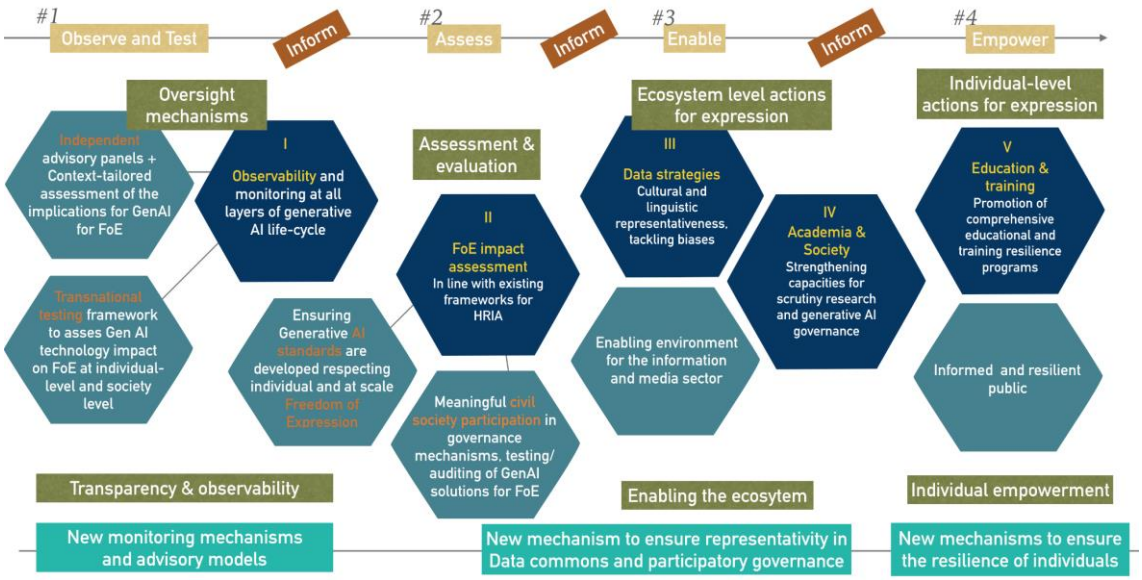
100. Les États membres, en collaboration avec la société civile, devraient soutenir les acteurs qui interviennent à tous les niveaux de la pile technologique de l'IA générative, afin de renforcer la transparence ; d'élargir les choix offerts aux utilisateurs ; inciter les acteurs économiques à adopter des pratiques commerciales qui respectent les droits fondamentaux ; et promouvoir une coordination internationale visant à partager les enseignements liés aux impacts sur la liberté d'expression. Une combinaison d'instruments réglementaires et non réglementaires peut être mobilisée pour corriger les dynamiques préjudiciables au sein de cet écosystème. Il peut s'agir, notamment, de codes de conduite sectoriels, d'avertissements émis par les autorités de régulation, de la publication d'évaluations des risques et des impacts, ainsi que d'indicateurs de performance pertinents pour les personnes qui souhaitent faire valoir leurs droits en matière de liberté d'expression.

Appendix 1



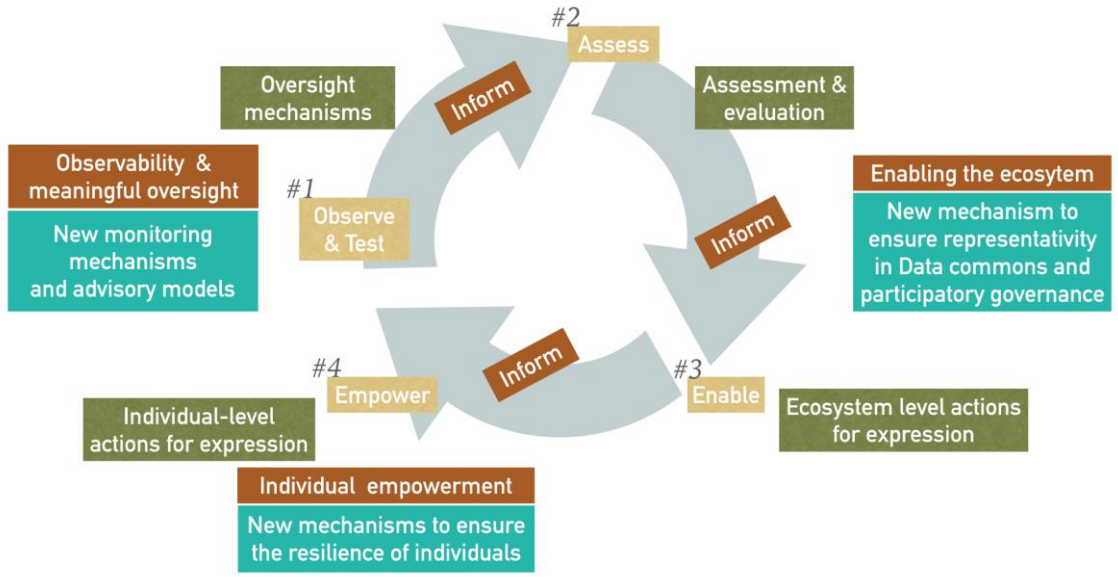
Anglais	Français
TECH	Couche fondamentale
TOOL	Couche fonctionnelle
PRODUCT	Couche applicative
Foundation Models and Core Technology	Modèles fondamentaux et technologies de base
Pre-training phase	Étape préalable à l'entraînement
Dataset	Jeu de données
Preprocessing	Prétraitement
Pre-training	Pré-entraînement
Data	Données
Data collection	Collecte de données
Data filtering	Filtrage de données
Model training	Entraînement du modèle
Training	Entraînement
Training Data processing and AI modeling	Traitement des données d'entraînement et modélisation de l'IA
System inherent risks for FOE	Risques inhérents au système pour la liberté d'expression
Content Policies and Technological Tuning	Politiques de contenu et réglage technologique
Model adaptation phase	Phase d'adaptation du modèle
Instructions	Instructions
Content alignment	Ajustement du contenu
Filtering	Fitrage
Curated data	Données sélectionnées
Instruction tuning	Ajustement par instructions
Align with human preferences	Alignement sur les préférences humaines
Guardrails and filters or systems prompts	Garde-fous, filtres ou invites système
System prompt	Invite système
Human labelling to fit to specific tasks	Étiquetage humain adapté à des tâches spécifiques
Reward to adapt	Récompense pour adapter le modèle
Output customization	Personnalisation des sorties

Refining and aligning models responses to adapt it to different tasks	Affinage et alignement des réponses du modèle pour les adapter à différentes tâches
Content moderation Risks for FOE	Risques liés à la modération du contenu pour la liberté d'expression
Application and Products Customization	Personnalisation des applications et produits
Product optimization phase	Phase d'optimisation du produit
Data augmentation	Augmentation des données
Agents	Agents
User customization	Personnalisation par l'utilisateur
Input customization	Personnalisation des entrées
Agentic workflows	Flux de tâches réalisées par des agents
Prompt assistance and user's custom features	Assistance à la rédaction de prompts et fonctionnalités personnalisées de l'utilisateur
Model/User Interaction Memory + User data for model final tuning	Mémoire des interactions modèle/utilisateur + Données utilisateur pour le réglage final du modèle
Data Retrieval Augmentation	Augmentation par récupération de données
Multi-models agents	Agents multi-modèles
Optimisation	Optimization
Optimizing and enhancing the inputs and outputs for end-use interactional purposes	Optimisation des entrées et sorties à des fins d'interaction avec l'utilisateur final
User facing Interactional risks for FOE	Risques interactionnels côté utilisateur pour la liberté d'expression



Anglais	Français
#1 Observe and Test	#1 Observer et tester
Oversight mechanisms	Mécanismes de contrôle
Independent advisory panels + Context-tailored assessment of the implications for GenAI for FoE	Comités consultatifs indépendants + évaluation contextuelle des conséquences de l'IA générative sur la liberté d'expression
Transnational testing framework to assess Gen AI technology impact on FoE at individual-level and society level	Cadre de test transnational pour évaluer l'impact des technologies d'IA générative sur la liberté d'expression au niveau individuel et sociétal
Transparency & observability	Transparence et traçabilité
New monitoring mechanisms and advisory models	Nouveaux mécanismes de suivi et modèles de conseil
#2 Assess	#2 Évaluer
Assessment & evaluation	Évaluation et analyse
FoE impact assessment in line with existing frameworks for HRIA	Évaluation de l'impact sur la liberté d'expression au regard des cadres existants d'évaluation des droits humains
Ensuring Generative AI standards are developed respecting individual and at scale Freedom of Expression	Faire en sorte que les normes d'IA générative respectent la liberté d'expression aux niveaux individuel et collective
Meaningful civil society participation in governance mechanisms, testing/auditing of GenAI solutions for FoE	Participation significative de la société civile aux mécanismes de gouvernance, de test et d'audit des solutions d'IA générative pour la liberté d'expression
#3 Enable	#3 Mettre en oeuvre
Ecosystem level actions for expression	Mesures prises au niveau de l'écosystème en faveur de l'expression humaine
Data strategies Cultural and linguistic representativeness, tackling biases	Stratégies de données Représentativité des cultures et des langues, Lutte contre les biais
Academia & Society Strengthening capacities for scrutiny research and generative AI governance	Renforcement des capacités du monde académique et de la société pour la recherche critique et la gouvernance de l'IA générative
Enabling the ecosystem	Mettre en oeuvre l'écosystème
New mechanism to ensure representativity in Data commons and participatory governance	Nouveaux mécanismes pour assurer la représentativité dans les jeux de données et gouvernance participative
#4 Empower	#4 Renforcer les capacités

Individual-level actions for expression	Actions individuelles en faveur de l'expression humaine
Education & training Promotion of comprehensive educational and training resilience programs	Éducation et formation Promotion de programmes complets d'éducation et de formation à la résilience
Informed and resilient public	Public informé et résilient
Individual empowerment	Renforcement des capacités individuelles
New mechanisms to ensure the resilience of individuals	Nouveaux mécanismes pour renforcer la résilience des individus



Anglais	Français
Observe & Test	Observer et tester
Assess	Évaluer
Assessment & evaluation	Évaluation et estimation
Oversight mechanisms	Mécanismes de contrôle
Observability & meaningful oversight	Traçabilité et contrôle
New monitoring mechanisms and advisory models	Nouveaux mécanismes de surveillance et modèles consultatifs
Inform	Informer
Enable	Mettre en oeuvre
Enabling the ecosystem	Mise en œuvre de l'écosystème
New mechanism to ensure representativity in Data commons and participatory governance	Nouveaux mécanismes pour assurer la représentativité dans les jeux de données et la gouvernance participative
Ecosystem level actions for expression	Mesures prises au niveau de l'écosystème en faveur de l'expression humaine
Empower	Renforcer les capacités
Individual-level actions for expression	Actions individuelles en faveur de l'expression humaine
Individual empowerment	Renforcer les capacités individuelles
New mechanisms to ensure the resilience of individuals	Nouveaux mécanismes pour renforcer la résilience des individus